Proceedings of

Vidya MCA Departmental Seminar (VMCADS – 2018)

22 – 23 November 2018

Editors

Dr V N Krishnachandran
Professor of Computer Applications

Aparna S Balan
Assistant Professor of Computer Applications

Manesh D
Assistant Professor of Computer Applications

Department of Computer Applications (MCA)
Vidya Academy of Science & Technology
(APJ Abul Kalam Technological University)
Thalakkottukara, Thrissur – 680501, India

Proceedings of Vidya MCA Departmental Seminar
(VMCADS – 2018)

Vidya Academy of Science & Technology
Thrissur – 680501, India

Distributed by MCA Department
Vidya Academy of Science & Technology
Thrissur – 680501, India

The book was typeset using the LaTeX document preparation system.

Cover design: Clinton Steephen

Price: Rs 0.00.

First printing: April 2019

Mission

Progress through Education

Vision

To seek, strive for and scale greater heights of quality education

Vidya Academy of Science & Technology

Contents

Preface

This volume contains the proceedings of the two-day departmental seminar organised by the Department of Computer Applications (MCA) of Vidya Academy of Science & Technology during 22 – 23 November 2018. The seminar was the culmination of a coursework (with course code RLMCA 352 Project and Viva Voce) to be completed by the MCA students of APJ Abdul Kalam Technological University during the Fifth Semester of the MCA programme.

The syllabus of the course RLMCA 352 Project and Viva Voce specifies the objective of the course as follows: "*To enable the students to gain knowledge in any of the technically relevant current topics on computer science/information technology/research, and acquire the confidence in presenting the topic and preparing a report.*"

As part of the course, each student is expected to undertake a detailed study on a technically relevant current topic in computer science/information technology under the supervision of a faculty member, by referring articles published in reputed journals/conference proceedings. Each student has to submit a seminar report, based on these papers; the report must not be reproduction of any original paper. The topic has to be presented taking a duration of 15 – 20 minutes. The report and slides for presentation has to be prepared using free typesetting software such as LaTeX.

In Vidya Academy of Science & Technology, the supervising teachers helped the students to identify the areas in which the students were to work and the teachers provided the students with some initial learning materials in the form of papers. After the initial reading of these materials, the students were asked to search for additional reading materials themselves. The students were required to study the papers and present a "study report/study paper" in a Departmental Seminar. The reports/papers collected in this volume are the study papers prepared by the students and presented in the Departmental Seminar. The Seminar was organised as a two-day event during 22 – 23 November 2018.

As part of the learning process, the students were also required to present the paper in the IEEE conference paper format. To facilitate this, the students were given a basic introduction to the LaTeX software and the `IEEEtran` document style.

In addition to gaining knowledge in any of the technically relevant current topics on computer science/information technology/research, the course also aimed at giving the students a hands on experience in preparing a conference/seminar paper. The expected learning outcomes included the following also:

- understanding of the structure of a research paper,
- awareness about the process of literature survey,
- basic knowledge about the accurate preparation of bibliography and their citations in the paper,

- exposure to the IEEE format for the preparation of conference/journal papers,
- introduction to the concepts of "Abstracts", "Keywords", and the like,
- experience in applying these concepts by actually preparing a paper, and
- methodology of presenting a multi-author paper in a seminar/conference.

The articles compiled in this Proceedings are not even moderately edited. The editors have only ensured that the basic learning outcomes outlined above have been met. However, the editors have tried to ensure that the titles of chapters, sections, etc., the abstract, figure and table captions, and the like are as per IEEE guidelines. The references have not been checked for accuracy and completion. The papers have not been edited for grammar, punctuation, spelling or style.[1]

The present work is only a record of the activities of the course referred to above and it is prepared only for private circulation. To the best of our understanding the authors of the papers have given proper attribution to ideas and material presented in the papers. If there are no attributions or improper attributions, it was unintentional. Hence the contents have not been subjected to plagiarism tests.

It is believed that the teachers as well the students doing the seminar course. There are still much scope for improvement. It is our hope that the future batches of students will have a stronger and wider learning experience from a similar seminar courses.

November 2018 Editors

[1] For different models of editing, see, for example "IEEE Editorial Style manual", [Online] Available: `https://www.ieee.org/documents/style_manual.pdf` (April 2017).

Proceedings of

Vidya MCA Departmental Seminar

(VMCADS – 2018)

Social Media: Uses and Gratification Perspective and Fake News Detection

Adeel Thaqib, Anju V R and Harikrishnan T S
Vidya Academy of Science & Technology
Thrissur - 680501, India

Siji K B
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: siji.k.b@vidyaacademy.ac.in)

***Abstract*—This paper presents importance of uses and gratifications theory to social media. By applying uses and gratifications theory, this paper will explore and discuss the uses and gratifications that consumer receive from using social media. This paper seeks to provide a better and more comprehensive understanding of why consumers use social media. This article provides a theoretical model to explicate the role of social media content in facilitating engagement behaviour within a social media context. Based on uses and gratifications theory, it provides a model for how an organisation can stimulate positively valenced engagement behaviour through social media and dissuade negatively valenced engagement behaviour in this forum. A typology of social media engagement behaviour is proposed and a series of hypotheses exploring the relationships between social media content and engagement behaviour are presented. Due to its rapid speed of information spread, wide user bases, and extreme mobility, Twitter is drawing attention as a potential emergency reporting tool under extreme events. This study explores the working dynamics of rumor mill by analyzing Twitter data of Haiti Earthquake 2010. For this analysis, two key variables of anxiety and informational certainty are employed from rumor theory, and their interactive dynamics are measured by both quantitative and qualitative methods. Our research finds that certain information with credible sources contribute to suppress the level of anxiety in Twitter community, which leads to rumor controlling and high information quality.**

***Index Terms*—Gratification theory, social media and engagement behaviour, social media fake news detection.**

I. Introduction

SOCIAL media is a critical area of interest for marketing scholars and practitioners. Recent research has shown that 88 percent of marketers are using social media and that they are spending over 60 billion annually on social media advertising. Successfully making contact with consumers via social media is predicted to show great returns for marketers in the coming years. The basic premise of uses and gratifications theory is that individuals seek out media that fulfill their needs and leads to ultimate gratification. Uses and gratifications theory has specific relevance to social media, but it has not been given prominence in the marketing and social media literature. Therefore, this paper seeks to apply uses and gratifications theory to help explain why consumers use social media. In particular, this research seeks to:

- Demonstrate the importance of uses and gratifications theory to social media.
- To apply uses and gratifications theory to social media.
- To identify the uses and gratifications that consumers receive from using social media.

By applying uses and gratifications theory, this research seeks to provide a better and more comprehensive understanding of why consumers use social media.The interactive properties of social media have transformed consumers from passive observers to active participants. Social trend shows that one type of social media does not replace another but, rather, becomes integrated into a bundle of media use that includes online and offline forms of communication.

Twitter has seemed to replace the legacy mainstream media. In 2008, before the Sichuan earthquake, which killed nearly 70,000 Chinese people, a tremor was tweeted by a local resident to the outside world. In January 2010, pictures of Haiti earthquake were first covered by Twitter and Facebook, which were later broadcast by CNN (Rosario 2010). Twitter reporting is so fast and real that, by the time mainstream media air those extreme events, they already become a trend word on Twitter search.The strength of Twitter lies not just on the speed of information spread. From the perspective of social media, another advantage is that, in a short period of time, Twitter users collectively cover major facets of disasters from multiple angles. As Twitter spreads multi-faceted eyewitness stories so rapidly, mainstream media companies follow Twitter to cite those tweets or pictures as their news sources. Despite many advantages, however, warnings have been raised about the information quality of Twitter.

During the Haiti earthquake, rumors circulated that UPS will “ship any package under 50 pounds to Haiti or “several airlines would take medical personnel to Haiti free of charge to help with earthquake relief. These turned out to be hearsay rather than eyewitness accounts, and subsequently refuted by

UPS and airline companies as false information. For this reason, Twitter is sometimes despised as a social media for propagation misinformation, rumors, and, in extreme case, propaganda. This criticism is somewhat acknowledged by Biz Stone, a co-founder of Twitter, when he said that credibility is key for social media. The context of this research is confined to extreme events such as natural disasters. The main thesis of this paper is largely to investigate social media in the extreme event scenario. The paper applies rumor theory to tweets posted during the Haiti Earthquake of 2010. The next section describes the theoretical framework by synthesizing two types of literatures: extreme events and rumor theory. Based on the result of data analysis, we subsequently carry out a semantic network analysis to further explore the tweet data.

II. Uses and Gratifications Theory

UGT is an approach to understanding why and how individuals actively seek out and use specific media to satisfy specific needs. Term gratifications to describe specific types or dimensions of satisfaction reported by audience members of daytime radio programmes. UGT addresses how individuals choose media that satisfies their needs, allowing one to realize gratifications such as knowledge enhancement, entertainment and relaxation, social interaction and reward or remuneration. While a UGT perspective has been applied in the context of television and electronic bulletins, the rapid growth of the Internet and social media platforms has created mediums in which a higher level of interactivity from users is required. The well-established theoretical perspective of UGT provides valuable insights into this new medium. As the underlying assumption of UGT is that users are actively involved in media usage, the theory has become increasingly relevant in studies of media channels that allow for consumer choice and interaction, such as social media.

In social media, a brands overt goal is toattract an audience by providing value, or gratification, through its content. Content must therefore be designed in a way which creates value for individual consumers to build a stronger level of engagement and facilitate value outcomes. Constructs based on the theoreticalunderpinnings of UGT, such as the need for social interaction, the need for entertainment,information seeking and sharing needs, and the desire for reward orremuneration have all been explored in recent literature that has investigated consumerchoices of online and social media. We posit that social media content can be categorized into four main groups based on its level of information, entertainment, remunerative and content.

A. Informational Content

80 percent of respondents reported using social media to seek out information. The informational construct of UGT represents the extent to which the social media content provides users with resourceful and helpful information. Whilst the importance of delivering information through advertisements has been recognized for traditional media, the role of informational content in the online, social domain has only recently received attention. Searching for and receiving information about a brand is one of the main gratifications of consumer participation in online brand communities. The desire to seek information directly from brands is a motivating factor for consumers to use social media.

B. Entertaining Content

64 percent of respondents reported that they used social media as a source of entertainment. The entertainment construct of UGT refers to the extent to which social media content is fun and entertaining to media users. Some of the entertainment activities reported were playing games, listening to music, and watching videos.The value of entertaining media is embedded in its ability to fulfill user needs for escapism, hedonistic pleasure, aesthetic enjoyment and emotional release. Others mentioned that they use social media for humor and comic relief. Some of their comments were listening to jokes, reading comments and stuff makes me laugh, and watching the crazies on Facebook, and how they display themselves, provides entertainment to me. Some respondents mentioned playing games regularly with friends via social networking platforms.

C. Remunerative Content

Consumers engage in social media use as they expect to gain some kind of reward such as an economic incentive, job-related benefit or personal wants. Social media content that offers a reward or remuneration includes monetary incentives, giveaways, prize drawings or monetary compensations. This type of content is expected to gratify users needs for remuneration and rewards within social media. Whilst managers often believe that social media content offering monetary incentives such as bonus points, prize draws or sharing product success are important, they are often mistaken.

D. Relational Content

Motivations for social media use include gaining a sense of belonging, connecting with friends, family and society, seeking support, and substituting real-life partnership. Users find the internet a comfortable place to reveal their feelings, share views and experiences, and to let their family and friends know about their latest information. Internet users expressed that through the online content generation process, they would have the opportunity to be recognized, publicize their expertise, learn more of the world, socialize with friends and be entertained. Socializing involves motivations such as gaining peer support, meeting interesting people, belonging to a community and staying in touch with friends.

III. Social Media Engagement Behavior

Social media is one of the more prevalent channels through which customers engage with a brand or firm, and businesses are recognising the need to engage where current and potential customers are paying most attention. Social media platforms provide users with an interactive avenue to create value and

engage with the firm. Users create social media content through their contributions, comments and likes. The notion of engagement has been studied in many fields, including psychology, education and management. A recent focus in marketing has centred on customer engagement with a brand. Social media engagement behaviours go beyond transactions, and may be specifically defined as a customers behavioural manifestations that have a social media focus, beyond purchase, resulting from motivational drivers.

A. Typologies of SMEB

In order to provide a deeper understanding of the behaviours consumers exhibit when they engage with social media, this paper proposes a typology of behaviours. The SMEB construct identifies and explicates the different types of engagement behaviours that users exhibit in social media platforms. It demonstrates that SMEB consists of seven distinct types; co-creation, positive contribution, consumption, dormancy, detachment, negative contribution and co-destruction. While co-creation, positive contribution, negative contribution and co-destruction represent active engagement behaviours that potentially impact on other social media users, consumption, dormancy and detachment are more passive and/or individualised forms of engagement. While the majority of current literature concerning customer engagement has focused on positively valenced engagement, the engagement concept can be extended to capture negatively valenced engagement. The construct of SMEB highlights the critical role of negatively valenced engagement behaviour within social media platforms. Negatively valenced SMEB includes detachment, negative contribution and co-destruction, exhibited through consumers unfavourable brand-related behaviors during interactions. Comparatively, positively valenced SMEB involves particular favourable or affirmative behavioural brand-related consumer dynamics. Positively valenced SMEB includes consumption, positive contribution and co-creation. In addition, the potential for an inactive, neutral state of engagement (termed dormancy) from the social media platform is recognised.

1) Co-creation: The highest possible level of positive, active SMEB whereby users initiate un-prompted, positive and active contributions to social media communities. Examples: Publishing a brand-related weblog, uploading brand-related video, audio, images, writing brand-related articles, reviews and testimonials.

2) Positive Contribution: A moderate level of positive, active SMEB whereby users make positive and active contributions to existing content on social media brand pages. Examples: Rating products, brands contributing to brand forums comment positively on posts, blogs, videos and pictures.

3) Consumption: The minimum level of positive, passive SMEB whereby users consume content without any form of active reciprocation or contribution. Example: Viewing brand-related video, listening to brand-related audio, viewing pictures and photos posted by the brand, reading brand posts, reading product/brand reviews.

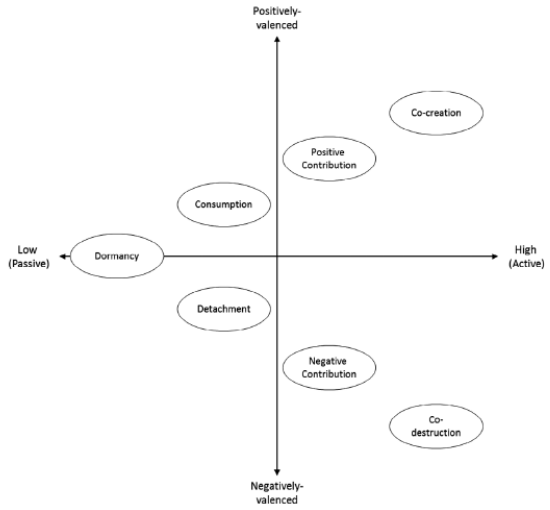

Fig. 1. Social media engagement typologies

4) Dormancy: A temporary state of inactive, passive SMEB by users who have previously interacted with the social media community. Examples: Brand-related content is delivered to the user but the user has no response.

5) Detachment: The minimum level of negative, active SMEB whereby users temporarily or permanently conclude their membership. Examples: Unliking or unfriending social media brand page, 'Unfollowing a brand on social media, terminating a subscription for further updates and content from the brand.

6) Negative Contribution: A moderate level of negative, active SMEB whereby users make negative contributions to existing content within the social media community. Examples: Conversing negatively on brand-related content, making negative contributions to brand forums, publicly rating products and brands negatively.

7) Co-destruction: The highest possible level of negative, active SMEB whereby users initiate unprompted, negative contributions to the social media community. Examples: Writing a public complaint, negative product reviews and testimonials, publishing a negative brand-related blog, initiating adverse social media brand pages for fellow community members to join.

IV. Social Media Content and Engagement Behaviour

Building on the previous discussion, we propose an integrative model of social media content and SMEB. The model suggests that social media content facilitates SMEB. There are seven discrete types of SMEB exhibited by users.

A. Engaging with Informational Content

Scholars have demonstrated that consumers engage when motivated by informational needs. This behavioural engagement manifests through actions such as clicking on links, staying on websites longer, reading details and threads and

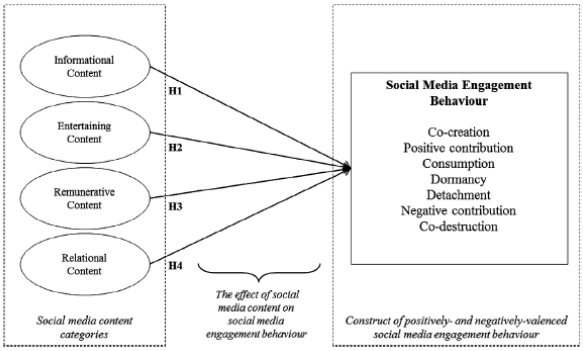

Fig. 2. Model of social media content and SMEB

using multimedia features. This denotes passive engagement with the brand, rather than active engagement in the form of commenting, or contributing to online communities. Informative content has been found to negatively impact levels of user engagement in the form of likes and comments, when compared to persuasive content such as emotional and philanthropic content. It is found that posts which contain information about the brand cause a lower level of engagement compared to entertaining content. H1: Informational content facilitates passive, positively-valenced social media engagement behaviour.

B. Engaging with Entertaining Content

Entertaining content can be found in messages which include small talk or banter, an attempt to gain trust and contain philanthropic content to appeal to a persons emotions. Cvijikj and Michahelles defined entertaining content as posts that did not refer to a brand or particular product, rather were written in the form of a teaser, slogan or word play. Entertaining content was a significant factor in increasing the number of likes, comments and shares made on social media content. H2: Entertaining content facilitates active, positively-valenced social media engagement behaviour.

C. Engaging with Remunerative Content

Remunerative social media content includes contests and sweepstakes and was found to be negatively related to the number of likes that a post received, but was a significant factor in predicting the number of comments. It is expected that a low level of behavioural engagement would occur as a result of a post containing a reward or offer, for example, consumption behaviour, rather than positive contribution or co-creation. It found that social media content which includes economic or remunerative information such as a product mention, price mention or deals and promotions has a negative impact on comments. H3: Remunerative content facilitates passive, positively-valenced social media engagement behaviour.

D. Engaging with Relational Content

Socialising refers to two-way, non-functional interactions through which consumers develop attitudes, norms and/or a common language. Customers with high social interaction motivations are more likely to engage in human-to-human interaction whether in an offline or online context. This interaction includes behaviours such as providing comments, feedback, personal information and participating in online discussion. These behaviours are reflective of the positive contribution and co-creation levels of the SMEB typology. H4: Relational content facilitates active, positively-valenced social media engagement behaviour.

V. Gratification for *Facebook* Use

Facebook is an SNS developed in 2004 by former Harvard undergraduate student Mark Zuckerberg, which allows users to add friends, send messages, and update personal profiles in order to notify friends and peers about themselves. *Facebook* users can also form and join virtual groups, develop applications,host content, and learn about each others interests,hobbies, and relationship statuses through users online profiles.Research into *Facebook* usage patterns suggests that *Facebook* is used and adopted primarily to maintain contact with offline connections rather than to develop new relationships. *Facebook* use was motivated primarily by social gratifications, which include maintaining existing social ties and being able to reconnect with friends from the past.

A. Gratification for Instant Messaging

IM is used primarily to fulfill needs including: affection, such as offering help and showing concern for others; entertainment, to have fun and to kill time; and relaxation, to get away from pressures and responsibilities. By contrast, *Facebook* is used primarily to keep in touch with old and current friends, to post/look at photographs, and to locate old friends. Finally, each medium provides different kinds of exposure situations that also affect the gratifications it provides. Even though all social media are characterized as interactive, there are differences between them in terms of the nature of interactions they support, which then leads to different types of gratifications.

VI. Method

A. Participants

Eighty-five participants were initially recruited from undergraduate courses in communications at a large, research intensive university in Canada. The sample was reduced to 77 after Facebook nonusers were removed.

B. Procedures

Participants were given a paper-and-pencil self-administered questionnaire. Twenty-one participants were recruited for the interviews through posters, which were displayed on bulletin boards across campus. Nineteen respondents participated in a face-to-face interview, and two respondents opted for an e-mail based interview. All interviews conducted face-to-face were recorded and transcribed with participants consent.

C. Results

Respondents are heavy users of Facebook: In the questionnaire, 82% reported logging into their Facebook account several times a day. The data showed also that students had been using Facebook for approximately one and a half years (M = 18.28, SD = 7.36). Five percent of respondents update their profile at least once a day, 22% update it at least once a week, 42% update it at least once a month, and 30% update it very rarely. The interview data show that respondents use Facebook extensively, logging into their accounts between two and five times per day.

VII. Gratification Obtained from *Facebook*

Eighty-five percent of questionnaire participants reported that their primary motivation for joining Facebook was that A friend suggested it. It is not surprising that friendship networks play an important role in the adoption of Facebook, considering that SNSs primary purpose is social connectivity.The second motivation chosen frequently by 49% was Everyone I know is on Facebook. Use of Facebook as a means for gettingaway from responsibilities and pressures and providing a formof entertainment.Two key gratifications were To kill time (M = 4.14, SD = 1.26) and Because itis entertaining (M = 4.04, SD = 1.13), showing how university students see Facebook as a diversion from other tasks and as a way to have fun. Factor 2, affection, comprises five items measuring how Facebook provides a venue for expressing concern and friendship toward others (eigenvalue = 4.53, variance explained = 19%).

Factor 1, pastime, comprises Leungs (2001) original entertainment, relaxation, and escape factors. Although key gratifications for both IM and Facebook are entertainment, relaxation, and escape, these seem to be more prominent in Facebook than IM.

Factor 2, affection, comprises five items measuring how Facebook provides a venue for expressing concern and friendship toward others (eigenvalue = 4.53, variance explained = 19%). This factor also completely overlapped with Leungs (2001) affection factor and showed how Facebook serves to reach out to others. IM is usually dyadic and allows for interactive conversations in real time that are somewhat comparable to face-to-face interactions. IM exchanges are linked to feelings of intimacy and the development of close ties.

Fashion was the third factor identified (eigenvalue = 2.71, variance explained = 11%) and consists of three items measuring the extent to which Facebook is fashionable. This factor overlapped with Leungs (2001) factor, even though the means were lower in the Facebook study than in Leungs study of gratifications of IM. Second, we expected that Facebook rather than IM would be used for gratifying fashion needs because IM is more private and less open than Facebook, where ones membership and profile could be seen much more as a display of trendiness.

Factor 4, sharing problems (eigenvalue = 1.64, variance explained = 7%) includes three items measuring the extent to which students use Facebook to talk to others about their concerns. Whereas two items loaded in the same manner as in Leungs (2001) study, the third item included in the factor was different in that it also revolved around sharing problems and not inclusion as in the IM study.

Social information is a factor that did not emerge as important in the IM gratifications study but is central in the analysis of the Facebook gratification structure. This is a new factor and consists of a single item To feel involved with whats going on with other people, which originally was part of the inclusion factor in Leungs (2001) study. Facebook provides not only more extensive information about users than IM but also qualitatively different information through the pictures section, the profile information, and the wall.

Where IM and Facebook intersect is on the inclusion dimension:Both Facebook and IM are seen as important tools for feeling involved with friends lives and keeping up-to-date with their activities. This creates a sense of membership in the peer community. A potential reason why IM users switch from IM to Facebook may be because the latter allows users to support much larger networks with less effort, whereas IM can quickly become overwhelming when the network size grows exponentially. In this way, Facebook supports larger volumes of exchanges with each exchange being much shorter and less involved and therefore easier to manage. Overall, we can conclude that sociability is a central gratification obtained from both forms of social media. However, the kinds of needs that each medium fulfills are different in nature and directly linked to their functionality.

VIII. Fake News Detection

"Fake news detection is defined as the task of categorizing news along a continuum of veracity, with an associated measure of certainty. Veracity is compromised by the occurrence of intentional deceptions. The paper provides a typology of several varieties of veracity assessment methods emerging from two major categories linguistic cue approaches (with machine learning), and network analysis approaches. We see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network-based behavioral data.

Fake news detection is defined as the prediction of the chances of a particular news article being intentionally deceptive. Tools aim to mimic certain filtering tasks which have, to this point, been the purview of journalists and other publishers of traditional news content. The proliferation of user-generated content, and Computer Mediated Communication (CMC) technologies such as blogs, Twitter, and other social media have the capacity of news delivery mechanisms on a mass scale yet much of the information is of questionable veracity. Structured datasets are easier to verify than non-structured (or semi-structured) data such as texts. When we know the language domain (e.g., insurance claims or health-related news) we can make better guesses about the nature and use of deception. Semi-structured non-domain specific web data come in many formats and demand flexible methods for veracity verification.

This paper provides researchers with a map of the current landscape of veracity (or deception) assessment methods, their major classes and goals, all with the aim of proposing a hybrid approach to system design.

1) Linguistic Approaches in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception.
2) Network Approaches in which network information, such as message meta-data or structured knowledge network queries can be harnessed to provide aggregate deception measures. Both forms typically incorporate machine learning techniques.

IX. Linguistic Approaches

Most liars use their language strategically to avoid being caught. In spite of the attempt to control what they are saying, language "leakage occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage. The goal in the linguistic approach is to look for such instances of leakage or, so called "predictive deception cues found in the content of a message.

A. Data Representation

Perhaps the simplest method of representing texts is the bag of words approach, which regards each word as a single, equally significant unit. In the bag of words approach, individual words or n-grams (multi-word) frequencies are aggregated and analyzed to reveal cues of deception. Further tagging of words into respective lexical cues for example, parts of speech or shallow syntax, affective dimensions, or location-based words are all ways of providing frequency sets to reveal linguistic cues of deception.

B. Deeper Syntax

Deeper language structures (syntax) have been analyzed to predict instances of deception. Deep syntax analysis is implemented through Probability Context Free Grammars (PCFG). Sentences are transformed to a set of rewrite rules (a parse tree) to describe syntax structure, for example noun and verb phrases, which are in turn rewritten by their syntactic constituent parts. The final set of rewrites produces a parse tree with a certain probability assigned. This method is used to distinguish rule categories (lexicalized, unlexicalized, parent nodes, etc.) for deception detection with 85-91 % accuracy.

C. Semantic Analysis

The intuition is that a deceptive writer with no experience with an event or object (e.g., never visited the hotel in question) may include contradictions or omission of facts present in profiles on similar topics. For product reviews, a writer of a truthful review is more likely to make similar comments about aspects of the product as other truthful reviewers. Extracted content from key words consists of attribute: descriptor pair.

X. Network Approaches

Innovative and varied, using network properties and behavior are ways to complement content-based approaches that rely on deceptive language and leakage cues to predict deception. As real-time content on current events is increasingly proliferated through micro-blogging applications such as Twitter, deception analysis tools are all the more important. The use of knowledge networks may represent a significant step towards scalable computational fact-checking methods. For certain data, false factual statements can represent a form of deception since they can be extracted and examined alongside findable statements about the known world. This approach leverages an existing body of collective human knowledge to assess the truth of new statements. The method depends on querying existing knowledge networks, or publicly available structured data, such as DBpedia ontology, or the Google Relation Extraction Corpus (GREC).

A. Characteristics of Extreme Events: Uncertainties

Large scale natural disasters are usually characterized by high consequence, low probability, ambiguity, and decision making pressure. The causes of natural disasters are uncontrollable and its future state is unpredictable, it easily renders infeasible the standard response and planning procedures. Therefore, it is essential for successful emergency response to anticipate the improvised assistance, ad-hoc communication endeavors, and adaptive collaboration among multiple agents such as firemen, police, volunteers, and governments with whom they have never collaborated before. Since the birth of Twitter, many web users have adopted Twitter as a tool for reporting their eyewitness accounts and relief activities during the natural disasters and relief operations.

B. Rumor Theory: Anxiety and Uncertain Information

Rumor is a form of collective behavior surrounding information and psychology. It is a collective transaction in which many people offer, evaluate, interpret information, and from which they predict something. When ignorance and ambiguity are removed, rumor disappears. The greaterthe anxiety, the more the content of rumor is important for the rumor recipient. Rrumor is a multiplicative function of anxiety and informational ambiguity. During the Haiti earthquake, many tweets, which were linked to reliable sources such as pictures, mainstream media, and organizations (such as Red Cross), contributed to reduce informational ambiguities and anxiety among networked citizens. According to rumor theory, during the large scale natural disasters, information quality (ambiguous or not), reliability of informational source, and level of anxiety among citizens are highly correlated. Following the logic of rumor theory, we posit that reduced anxiety and enhanced informational certainty in Twitter cyberspace can suppress rumors, and that is the necessary, although not sufficient, condition for emergence of problem solving discourse.

XI. METHODS

To collect data, we used #haitiearthquake as a search keyword. In addition, lots of tweets returned by the #Haiti keyword included posts which are not topically relevant to the Haiti earthquake. After trying several searches with various hash-keywords, we found that many users posted incident reports by attaching multiple hash-keywords such as #HaitiEarthquake, #HaitiQuake, or #HaitiHelp within their messages. As these three hash-keywords seemed to be the most frequently used and topically relevant ones made in English language.

A. Coding Scheme

To identify the content types of tweet posts, we used the Rumor Interaction Analysis Systems (RIAS) as our coding scheme, which was originally developed by Bordia. The updated RIAS includes 14 categories of communication modes: prudent, apprehensive, authenticating, interrogatory, providing information, belief, disbelief, sense making, directive, sarcastic, wish, personal involvement, digressive, and uncodable. From these 14 categories, we dropped 7 categories of digressive, personal involvement, wish, sarcastic, apprehensive, providing information, and sense making.

XII. CONCLUSION

The application of uses and gratifications theory to social media helps explain the many and varied reasons why consumers use and like social media. The findings from the in-depth interviews provide a very rich and comprehensive understanding of why consumers utilize social media. These findings can help businesses to more effectively market to and communicate with its existing and potential customers. Through the examination of the role of social media content using UGT, the paper contributes to a deeper understanding of engagement behaviour within social media platforms. We explored the influence of informational content, entertaining content, relational content and remunerative content on positively and negatively valenced engagement behaviour. The current study identifies ten uses and gratifications for using social media. The ten uses and gratifications found in this study are social interaction (88 percent), information seeking (80 percent), pass time (76 percent), entertainment (64 percent), relaxation (60 percent), communicatory utility (56 percent), expression of opinions (56 percent), convenience utility (52 percent), information sharing (40 percent), and surveillance and watching of others (20 percent).

Linguistic and network-based approaches have shown high accuracy results in classification tasks within limited domains. This discussion drafts a basic typology of methods available for further refinement and evaluation, and provides a basis for the design of a comprehensive fake news detection tool. Linguistic processing should be built on multiple layers from word/lexical analysis to highest discourse-level analysis for maximum performance. As a viable alternative to strictly content-based approaches, network behavior should be combined to incorporate the trust dimension by identifying credible sources. Contributions in the form of publicly available gold standard datasets should be in linked data format to assist in up-to-date fact checking. Rumors are a collective behavior surrounding information and based on the psychology of humans. Rumors involve an activity of creating, exchanging, and evaluating information at the collective level (Shibutani 1966). This paper analyzed Twitter data of Haiti earthquake 2010 through the lens of rumor theory. Results of both quantitative and qualitative analysis validate that anxiety and informational ambiguity are key variables to understand abnormal communication patterns under extreme events. Our finding confirms that reliable information with credible sources can contribute to reduce anxiety, suppressing groundless rumor.

REFERENCES

[1] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen, "Automatic Deception Detection: Methods for Finding Fake News", *Language and Information Technology Research Lab*, Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, Canada.

[2] Anabel Quan-Haase and Alyson L. Young, "Uses and Gratifications of Social Media: A Comparison of Facebook and Instant Messaging", *Bulletin of Science Technology & Society*, 2010, 30:350.

[3] Anita Whiting and David Williams, "Why people use social media: a uses and gratifications approach", College of Business, Clayton State University, Morrow, Georgia, USA; Department of Marketing, Berry College, Mount Berry, Georgia, USA

[4] Rebecca Dolan, Jodie Conduit, John Fahy & Steve Goodman, "Social media engagement behaviour: a uses and gratifications perspective", *Journal of strategic marketing*, Vol. 24.2016, 3/4, p. 261-277, Taylor & Francis.

[5] Onook Oh, Kyounghee Hazel Kwon, H. Raghav Rao, "An exploration of social media in extreme events: Rumour theory and twitter during the Haiti earthquake 2010", 2010, *ICIS 2010 Proceedings*, 231.

A Survey on Web Structure Mining

Aiswarya M A, Ginex Nedumparambil, Hamsheena K V and Haritha P M
Vidya Academy of Science & Technology
Thrissur - 680501, India

Aparna S Balan
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: aparna@vidyaacademy.ac.in)

***Abstract*—Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In this paper, we provide an introduction of Web mining as well as a survey of the Web Structure mining by taking papers from 2008 to 2015.**

I. Introduction

TODAY, the World Wide Web is the popular and interactive medium to disseminate information. The Web is huge, diverse and dynamic. The Web contains vast amount of information and provides an access to it at any place at any time. Most of the people use internet for retrieving information. But most of the time, they get lots of insignificant and irrelevant document even after navigating several links.

For retrieving information from the Web, Web mining techniques are used. [1] Web mining can be considered as the applications of the general data mining techniques to the Web. However, the intrinsic properties of the Web make us have to tailor and extend the traditional methodologies considerably.

- Firstly, even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together.
- Secondly, Web pages are semi-structured, in order for easy processing; documents should be extracted and represented into some format.
- Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough.

Even though the difficulties above, the Web also provides other ways to support mining, for example, the links among Web pages are important resource to be used. Besides the challenge to find relevant information, users could also find other difficulties when interacting with the Web such as the degree of quality of the information found, the creation of new knowledge out of the information available on the Web, personalization of the information found and learning about other users. Web mining techniques could be applied to solve, partially or completely, the above cited problems. However, Web mining techniques are not the only tools to solve those problems.

II. Web Mining

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. To clarify the confusion of what forms Web mining. Kosala and Blockeel had suggested a decomposition of Web mining in the following tasks:

1) *Resource finding*:
   The task of retrieving intended Web documents.
2) *Information selection and pre-processing*:
   Automatically selecting and pre-processing specific information from retrieved Web resources.
3) *Generalization*:
   Automatically discovers general patterns at individual Web sites as well as across multiple sites.
4) *Analysis*:
   Validation and/or interpretation of the mined patterns.

In general, Web mining tasks can be classified into three categories :

1) Web content mining
2) Web structure mining
3) Web usage mining

III. Web Structure Mining

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of

research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [8], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1) *Link-based Classification*:
   Link-based classification is the most recent upgrade of a classic data mining task to linked domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.
2) *Link-based Cluster Analysis*:
   The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.
3) *Link Type*:
   There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities,or predicting the purpose of a link.
4) *Link Strength*
   Links could be associated with weights.
5) *Link Cardinality*:
   The main task here is to predict the number of links between objects.

IV. Web Search and Hyperlinks

Traditional information retrieval systems and search engines first extract relevant documents to users based on content similarity query entered and indexed pages. In the late 1990s, it was concluded that the methods used are based on content alone are not sufficient due to the large volume of information available on the Internet. When applying a query using a search engine the page numbers results relevant to this query is very high. Thus, to meet the satisfaction of users, search engines must choose the first 30-40 pages results of relevant query. Thus, there are used hyperlinks that connect pages together. In 1998, there were created two very important algorithms based on hyperlinks, PageRank and HITS. Both algorithms, PageRank (Palau, et al., 2004) and HITS (Kleinberg, 1998), draw their origin from social network analysis. They use the hyperlinks structure of the web pages to give ranks according to the degree of prestige or authority. Page Rank algorithm was created in 1998 by Sergey Brin and Larry Page. Based on this algorithm it works the most successful Internet search engine, Google.

Page Rank is rooted in social network analysis, it basically provide a ranking of each web page depending on how many links from other sites leading to that page. The key idea is to use the probability that a page is visited by a random surfer on the Web as an important factor for ranking search results. This probability is approximated by the so-called page rank, which is again computed iteratively. The popularity (or prestige) of a web page can be measured in terms of how often an average web user visits it. To estimate this, we may use the metaphor of the random web surfer, who clicks on hyperlinks at random with uniform probability and thus implements the random walk on the web graph. Assume that page u links to Nu web pages and page v is one of them. Then once the web surfer is at page u, the probability of visiting page v will be 1/Nu. This intuition suggests a more sophisticated scheme of propagation of prestige through the web links also involving the out-degree of the nodes. The idea is that the amount of prestige that page v receives from page u is 1/Nu from the prestige of u. This is also the idea behind the web page ranking algorithm PageRank

V. Web Page Ranking Algorithms

The search engines on the Web need to be more efficient because there are extremely large number of Web pages as well queries submitted to the search engines. Web mining techniques are employed by the search engines to extract relevant documents from the web database and provide the necessary information to the users. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the analysis of the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document. There are number of algorithms proposed based on link analysis. Three important algorithms, such as PageRank, Weighted PageRank and HITS (Hyper-link Induced Topic Search) are discussed below.

A. Page Rank Model

L. Page and S. Brin proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach, Brin and Page extend the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as : "We assume page A has pages T1...Tn which point to it (that is, are citations)." The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also $C(A)$

is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

$$\mathrm{PR}\,(A) = (1-d) + d\left(\frac{\mathrm{PR}\,(T1)}{C(T1)} + \cdots + \frac{\mathrm{PR}\,(Tn)}{C(Tn)}\right).$$

Note that the rank of a page is divided evenly among its outlinks to contribute to the ranks of the pages they point to. The equation is recursive, but starting with any set of ranks and iterating the computation until it converges may compute it. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages.

They use the hyperlinks structure of the web pages to give ranks according to the degree of prestige or authority. Page Rank algorithm was created in 1998 by SergeyBrin and LarryPage. Based on this algorithm it works the most successful Internet search engine, Google. Page Rank is rooted in social network analysis.

The key idea is to use the probability that a page is visited by a random surfer on the Web as an important factor for ranking search results .This probability is approximated by the so-called page rank , which is again computed iteratively .The popularity (or prestige) of a webpage can be measured in terms of how often an average web user visits it .To estimate this,we may use the metaphor of therandom web surfer, who clicks on hyperlinks at random with uniform probability and thus implements the random walk on the web graph.

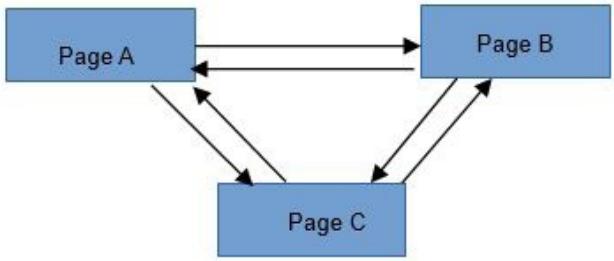

Fig. 1. Incoming and outgoing links

Result analysis: You can notice that PageRank of C is higher than PageRank of B and A. It is because Page C has 2 incoming links and 2 outgoing links as shown in Fig. 1. Page B has 2 incoming links and 1 outgoing link. Page A has the lowest PageRank because Page A has only one incoming link and 2 outgoing links. So the link analysis becomes very important in the PageRank. The PageRank gets converged to a reasonable tolerance.

Note that the rank of a page is divided evenly among its outlinks to contribute to the ranks of the pages they point to. The equation is recursive, but starting with any set of ranks and iterating the computation until it converges may compute it. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages [15].

The PageRank forms a probability distribution curve over the Web pages. PageRank can be mathematically computed using normalized eigenvector equations by iteratively processing values until all converge to a single nonrepeating value.

Advantages

- It computes the rank of web pages at crawling time, hence the response to user query is quick.
- It is less susceptible to localized links as it uses the entire web graph to generate page ranks rather than a small subset

Disadvantages

- It leads to spider traps if a group of pages has no outlinks to another external group of pages.
- Dead ends and circular references will reduce the front pages PageRank.
- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included.
- In Internet, available data is huge and the algorithm is not fast enough.
- It should support personalized search that personal specifications should be met by the search result.

B. Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani proposed a Weighted PageRank algorithm which is an extension of the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than Dividing the rank value of page evenly among its outgoing linked pages, each outgoing link gets a value proportional to its importance.

The more popular web pages are, the more linkages that other webpages tend to have to them or are linked to by them. The proposed extended Page Rank algorithma Weighted Page Rank Algorithmassigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W^{in}(v,u)$ and $W^{out}(v,u)$, respectively.

$W^{\mathrm{in}}(v,u)$ is the weight of $link(v,u)$ calculated based on the number of in-links of page u and the number of in-links of all reference pages of page v.

$$w^{\mathrm{in}}(v,u) = \frac{I_u}{\sum_{p\in R(v)} I_p}$$

where I_u and I_p represent the number of in-links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v. $W^{\mathrm{out}}(v,u)$ is the weight of $link(v,u)$ calculated based on the number of out-links of page u and the number of out-links of all reference pages of page v.

$$w^{\mathrm{out}}(v,u) = \frac{O_u}{\sum_{p\in R(v)} O_p}$$

where O_u and O_p represent the number of out-links of page $u$and page p, respectively. $R(v)$ denotes the reference page list of page v.

To evaluate the WPR algorithm, we implemented WPR and the standard PageRank algorithms to compare their results. The simulation studies we have carried out in this work consist of six major activities:

Experiments:

1) Finding a web site: Finding a web site with rich hyperlinks is necessary because the standard PageRank and the WPR algorithms rely on the web structure. After comparing the structures of several web sites, the website of Saint Thomas University, in Fredericton, has been chosen.
2) Building a web map: There is no web map available for this website. A free spider softwareJSpideris used to generate the required web map.
3) Finding the root set: A set of pages relevant to a given query is retrieved using the IR search engine embedded in the web site. This set of pages is called the root set.
4) Finding the base set: A base set is created by expanding the root set with pages that directly point to or are pointed to by the pages in the root set.
5) Applying algorithms: The Standard PageRank and the WPR algorithms are applied to the base set.
6) Evaluating the results: The algorithms are evaluated by comparing their results.

C. HITS Concept

In HITS concept, Kleinberg [14] identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg [14], Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

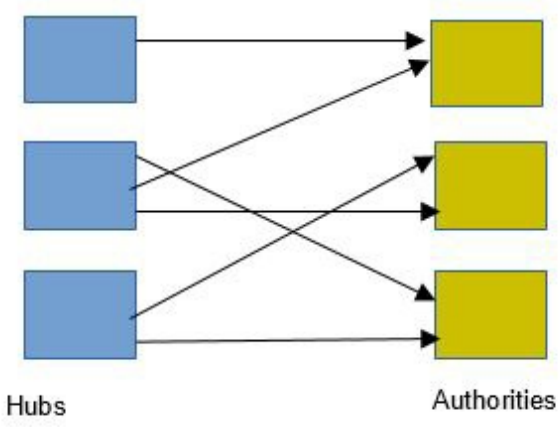

Fig. 2. Hubs and Authorities

The Hits algorithm has two steps:

1) Sampling Step - in this step a set of relevant pages for the given query are collected.
2) Iterative Step - in this step Hubs and Authorities are found using the output of sampling step.

The algorithm performs a series of iterations, each consisting of two basic steps:

- Authority update: Update each node's authority score to be equal to the sum of the hub scores of each node that points to it. That is, a node is given a high authority score by being linked from pages that are recognized as Hubs for information.
- Hub update: Update each node's hub score to be equal to the sum of the authority scores of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

An important difference between PageRank and HITS is the way that page scores are propagated in the web graph. According to this algorithm first step is to collect the root set. That root set hits from the search engine. Then the next step is to construct the base set that includes the entire page that points to that root set. The size should be in between 1000-5000. Third step is to construct the focused graph that includes graph structure of the base set. It deletes the intrinsic link, (the link between the same domains). Then it iteratively computes the hub and authority scores. In HITS concept, he identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to him, A good hub is a page that points to many good authorities; a Good authority is a page that is pointed to by many good hubs. Although HITS provides good search results for a wide range of queries.

The following are the constraints of HITS algorithm

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- Automatically generated links: HITS gives equal importance for automatically generated links which may not produce relevant topics for the user query.

VI. Comparison

In PageRank and Weighted PageRank algorithm WSM technique is used for mining,where as in HITS algorithm both WSM and WCM is used. Working of PageRank and Weighted PageRank is based on computing scores at index time and the results are sorted on the importance of pages. HITS algorithm computes scores of n highly relevant pages. Complexity of these three algorithm is $O(logN)$. Limitation of PageRank and Weighted PageRank algorithms are query independent. Topic drift and efficiency problem is the drawback of HITS algorithm

VII. Conclusion

This study covers the basics of Web mining. The importance of the Web structure mining in Information retrieval is explained.The main purpose of this study is to explore the hyperlink structure and understand the Web graph in a simple way. The PageRank computation results shows that the incoming links and the outgoing links play an important role in ranking of Web pages using link analysis.. This study also focuses on the important algorithms used for hyperlink analysis, explore those algorithms and compare them. This study is done basically to explore the link structure algorithms for ranking and compare those algorithms.The further work on this area will be problems facing PageRank algorithm and how to handle those problem

References

[1] M. G. da Gomes Jr. and Z. Gong, "Web Structure Mining: An Introduction", *Proceedings of the IEEE International Conference on Information Acquisition*, 2005.

[2] R. Kosala, H. Blockeel, "Web Mining Research: A Survey, *SIGKDD Explorations*, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[3] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, 46(5):604632, September 1999.

[4] R. Kosala and H. Blockeel, "Web mining research: A survey", *ACM SIGKDD Explorations*, 2(1):115, 2000.

[5] S. Madria, S. S. Bhowmick, W. K. Ng, "Re-search issues in web data mining", *Proceedings of the Conference on Data Warehousing and Knowledge Discovery*, pp. 303319, 1999.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[7] P. Ravi Kumar and Ashutosh Kumar Singh, "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", Department of Electrical and Computer Engineering, Curtin University of Technology, Sarawak Campus, Miri, Malaysia,

Image Segmentation: Techniques and Applications in Medical Field

Aiswarya K L, Manju M Krishnadas and Smera P R
Vidya Academy of Science & Technology
Thrissur- 680501, India

Reji C Joy
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: reji.c.j@vidyaacademy.ac.in)

***Abstract*—The Image segmentation is referred to as one of the most important processes of image processing. Image segmentation is the technique of dividing or partitioning an image into parts, called segments. It is mostly useful for applications like image compression or object recognition, because for these types of applications, it is inefficient to process the whole image. Medical image segmentation plays a crucial role in delineation of regions of interest under study. It is essential in almost any medical imaging applications and is an essential step towards automated disease state detection in diagnostic imaging.Medical images are at the core of medical science and an enormous source of information that need to be utilized.Image processing techniques with regards to biomedical images are generally either used for the retrieval of images or for analysis and modification of images.**

***Index Terms*—digital image,segmentation,pixel,clustering.**

I. Introduction

IMAGE processing is now routinely used by a wide range of individuals who have access to digital cameras and computers. With a minimum investment, one can readily enhance contrast, detect edges, quantify intensity, and apply a variety of mathematical operations to images. Although these techniques can be extremely powerful, the average user often digitally manipulates images with abandon, seldom understanding the most basic principles behind the simplest image-processing routines. Digital image processing is the use of computer algorithms to perform image processing on digital images. Image segmentation is an important and challenging process of image processing. Image segmentation technique is used to partition an image into meaningful parts having similar features and properties. The basic applications of image segmentation are: Content-based image retrieval, Medical imaging, Object detection and Recognition Tasks, Automatic traffic control systems and Video surveillance, etc. The image segmentation can be classified into two basic types: Local segmentation (concerned with specific part or region of image) and Global segmentation (concerned with segmenting the whole image, consisting of large number of pixels).

II. Image Segmentation

Segmentation is one of the key problems in image processing. Image segmentation is the process that subdivides an image into its constituent parts or objects. The level to which this subdivision is carried out depends on the problem being solved, i.e., the segmentation should stop when the objects of interest in an application have been isolated e.g., in autonomous air-to-ground target acquisition, suppose our interest lies in identifying vehicles on a road, the first step is to segment the road from the image and then to segment the contents of the road down to potential vehicles. Image thresholding techniques are used for image segmentation. After thresholding a binary image is formed where all object pixels have one gray level and all background pixels have another - generally the object pixels are 'black' and the background is 'white'. The best threshold is the one that selects all the object pixels and maps them to 'black'. Various approaches for the automatic selection of the threshold have been proposed. Thresholding can be defined as mapping of the gray scale into the binary set 0, 1}:

III. Image Segmentation Techniques

A. Thresholding Method

Thresholding methods are the simplest methods for image segmentation. These methods divide the image pixels with respect to their intensity level. These methods are used over images having lighter objects than background. The selection of these methods can be manual or automatic i.e. can be based on prior knowledge or information of image features. There are basically three types of thresholding:

1) Global Thresholding
2) Variable Thresholding
3) Multiple Thresholding

The values of thresholds can be computed with the help of the peaks of the image histograms. Simple algorithms can also be generated to compute these.

B. Edge Based Segmentation Method

The edge detection techniques are well developed techniques of image processing on their own. The edge based segmentation methods are based on the rapid change of intensity value in an image because a single intensity value does not provide good information about edges. Edge detection techniques locate the edges where either the first derivative of intensity is greater than a particular threshold or the second derivative has zero crossings. In edge based segmentation methods, first of all the edges are detected and then are connected together to form the object boundaries to segment the required regions.

C. Region Based Segmentation Method

The region based segmentation methods are the methods that segments the image into various regions having similar characteristics.There are two basic techniques based on this method:

1) Region growing methods: The region growing based segmentation methods are the methods that segments the image into various regions based on the growing of seeds (initial pixels).
2) Region splitting and merging methods: The region splitting and merging based segmentation methods uses two basic techniques i.e. splitting and merging for segmenting an image into various regions.

D. Clustering Based Segmentation Method

The clustering based techniques are the techniques, which segment the image into clusters having pixels with similar characteristics. Data clustering is the method that divides the data elements into clusters such that elements in same cluster are more similar to each other than others.

1) Hard Clustering: Hard clustering is a simple clustering technique that divides the image into set of clusters such that one pixel can only belong to only one cluster. In other words it can be said that each pixel can belong to exactly one cluster.
2) Soft clustering: The soft clustering is more natural type of clustering because in real life exact division is not possible due to the presence of noise.

E. Watershed Based Methods

The watershed based methods uses the concept of topological interpretation. In this the intensity represents the basins having hole in its minima from where the water spills. When water reaches the border of basin the adjacent basins are merged together. To maintain separation between basins dams are required and are the borders of region of segmentation. These dams are constructed using dilation. The watershed methods consider the gradient of image as topographic surface. The pixels having more gradient are represented as boundaries which are continuous.

F. Partial Differential Equation Based Segmentation Method

The partial differential equation based methods are the fast methods of segmentation. These are appropriate for time critical applications. The results of the PDE method is blurred edges and boundaries that can be shifted by using close operators. The fourth order PDE method is used to reduce the noise from image and the second order PDE method is used to better detect the edges and boundaries.

G. Artificial Neural Network Based Segmentation Method

The artificial neural network based segmentation methods simulate the learning strategies of human brain for the purpose of decision making. Now days this method is mostly used for the segmentation of medical images. It is used to separate the required image from background. A neural network is made of large number of connected nodes and each connection has a particular weight. This method is independent of PDE. In this the problem is converted to issues which are solved using neural network.

IV. Applications in Medical Field

A. Coma Patient Monitoring System with Feedback

To provide more convenient and comprehensive medical monitoring in big hospitals since it is tough job for medical personnel to monitor each patient for 24 hours. The latest development in patient monitoring system can be used in intensive care unit (ICU), critical care unit (CCU), and emergency rooms of hospital. During treatment, the patient monitor is continuously monitoring the coma patient to transmit the important information. Also in the emergency cases, doctor are able to monitor patient condition efficiently to reduce time consumption, thus it provides more effective healthcare system. This project investigates about the effects seen in the patient using image processing based coma patient monitoring system with feedback which is a very advanced project related to physical changes in body movement of the patient and It also provide a system for monitoring patient heartbeat rate and gives warning in form of alarm & LCD Display. It also passes a SMS to a person sitting at the distant place if there exists any movement in any body part of the patient.

Viola Jones algorithms use for real time body movement. There are three key contribution fast and accurate detection; the integral image for feature computation, ad boost for feature selection and an intentional cascade for efficient computational resource allocation. This system will consist of an input unit Digital camera to study the movement in any part of the body of patient and measure heartbeat of patient using heartbeat sensor. For indication of warning we will use by sending SMS through the GSM mobile to a distant person and display warning on LCD display. Moreover the warning will be deactivated manually rather than automatically. So for this purpose a deactivation switch will be used to deactivate warning.

Materials and Methods

- GSM Module
- Arduino
- Heartbeat Sensor
- Block Diagram
- Flow Chart
- Input Stage
- Output Stage

B. Preprocessing and Segmentation Techniques on Cardiac Medical Images

The heart is a vital organ of the human circulatory system. Proper functioning of the heart is essential to prevent Cardio Vascular Diseases (CVD). Lack of cardio vascular exercise, sedentary and stressed lifestyle, unhealthy diet, diabetes and genetic factors, all contribute to the development of cardio vascular disorders. Generally, cardiac examination involves assessing a combination of the four following physiological measures: cardiac structure, function, perfusion and myocardial viability. This topic has been investigating for decades and it is still an active research field. The main challenges include wide shape variability of heart between different cardiac cycles and between different patients, weak edges between epicardium and heart fat or soft tissues.

Materials and Methods

1) Cardiac Image Pre-processing Techniques: Since most of the real life data is noisy, inconsistent and incomplete, so preprocessing becomes necessary. Image preprocessing is one of the Preliminary steps which are highly required to ensure the high accuracy of the subsequent steps.
2) Cardiac Image Segmentation Techniques: An image in computer vision system could be defined as a two dimensional or three dimensional matrices of pixels, where each pixel corresponds to a definite intensity value. In medical imaging, these intensities could be radiation absorbed during x-ray, or acoustic pressure in ultrasonography, or radio frequency signals in MRI etc. Image segmentation is a procedure in which an image is divided into regions of some homogeneous characteristics like grey scale value, colour or texture. Cardiac image segmentation methods could be broadly categorized into following:

- Histogram Based Methods
- Statistical Model Based Methods
- Region Based Methods
- Graph Based Methods
- Deformable Model Based Methods
- Atlas Based Methods

C. Lung Cancer Detection using Digital Image Processing on CT Scan Images

Lung cancer mortality rate is the highest among all other types of cancer. It is one of the most serious cancers in the world, with the survival rate very less after the diagnosis. Survival from lung cancer is directly related to its growth at its detection time. The earlier the detection is, the higher the chances of successful treatment are. An estimated 85% of lung Cancer cases in males and 75% in females are caused by cigarette smoking . There are many techniques to diagnose lung cancer, like Chest Radiography (x-ray), Computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) but, most of these techniques are costly and time consuming. And most of these techniques are detecting the lung cancer in its advanced stages. Hence, there is a great need of a new technology to diagnose the lung cancer in its early stages. Image processing techniques provide a good class tool for cultivating the manual analysis.

Methodology

1) Image Enhancement: There are two types of enhancement technique, Special domain and Frequency domain. Due to enhancement we improve the quality of images, for human viewer or to provide better input to image processing technique. We use histogram equalization technique for enhancement.
2) Image Segmentation: Segmentation is nothing but the partition of image. Segmentation is typically use to detect object and boundaries of an image. We use watershed segmentation technique. Watershed segmentation extract seeds indication the presence of object or background at ct scan image. The marker location is then set to be regional minima typically gradient of the original input image and the watershed algorithm is applied.
3) Feature Extraction: It is important stage in image processing technique. It detects desired portion or shape of an image. For the classification purpose we need the features as like area, perimeter, roundness, eccentricity etc are used.

V. Conclusion

Image segmentation is one of the most important requirements of image processing in which an image is partitioned into a set of objects and backgrounds. Segmentation plays vital role in analyzing an image automatically. The main objective of segmentation is to trace certain objects of interest by ignoring the effect of light, noise and texture on them. Accurate segmentation of medical images is a key step in contouring during radiotherapy planning. Computed topography (CT) and Magnetic resonance (MR) imaging are the most widely used radiographic techniques in diagnosis, clinical studies and treatment planning. The goals of computer-aided diagnosis (CAD) are:

- To automate the process so that large number of cases can be handled with the same accuracy i.e. the results are not affected as a result of fatigue, data overload or missing manual steps.
- To achieve fast and accurate results. Very high-speed computers are, now, available at modest costs, speeding up computer-based processing in the medical field.
- To support faster communication, wherein patient care can be extended to remote areas using information technology.

The techniques available for segmentation of medical images are specific to application, imaging modality and type of body part to be studied. There is no universal algorithm for segmentation of every medical image. Each imaging system has its own specific limitations

REFERENCES

[1] aurav Kumar, Pradeep Kumar Bhatia A Detailed Review of Feature Extraction in Image Processing Systems 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

[2] Rafael C. Gonzalez and Richard E. Woods, Digital Image Processing, 2nd ed., Beijing: Publishing House of Electronics Industry, 2007.

[3] T. Shraddha, K. Krishna, B.K.Singh and R. P. Singh, Image Segmentation: A Review, International Journal of Computer Science and Management Research Vol. 1 Issue. 4 November 2012.

[4] M. R. Khokher, A. Ghafoor and A. M. Siddiqui, Image segmentation using multilevel graph cuts and graph development using fuzzy rule-based system, IET image processing, 2012.

[5] V. Dey, Y. Zhang and M. Zhong, a review on image segmentation techniques with Remote sensing perspective, ISPRS, Vienna, Austria, July 2010.

[6] S. Inderpal and K. Dinesh, A Review on Different Image Segmentation Techniques, IJAR, Vol. 4, April, 2014.

[7] Suzuki K., false-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs,academic radiology, volume 13, february 2005

[8] Gaurav Kumar, Pradeep Kumar BhatiaA Detailed Review of Feature Extraction in Image Processing Systems 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

[9] H.C. van Assen, M.G. Danilouchkine, F. Behloul, H.J. Lamb, R.J.vanderGeest, J.H.C. Reiber, and B.P.F. Lelieveldt,Cardiac LV Segmentation Using a 3D Active Shape Model Driven by Fuzzy Inference,MICCAI 2003

[10] Olivier Ecabert, Jochen Peters, Hauke Schramm, Cristian Lorenz, Jens von Berg, Matthew J. Walker,ManiVembar, Mark E. Olszewski, Krishna Subramanyan, Guy Lavi, and JrgenWeese Automatic Model-Based Segmentation of the Heart in CT Images,IEEE

[11] H.C. van Assen, M.G. Danilouchkine, F. Behloul, H.J. Lamb, R.J.vander Geest, J.H.C. Reiber, and B.P.F. Lelieveldt,Cardiac LVSegmentation Using a 3D Active Shape Model Driven by FuzzyInference,MICCAI 2003.

Introduction to Ajax Technology

Aiswarya B, Nikhil M A and Tison C Sunny
Vidya Academy of Science & Technology
Thrissur 680501, India

Salkala K S
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur 680501, India
(email: salkala@vidyaacademy.ac.in)

***Abstract*—Asynchronous JavaScript and XML (Ajax) is a Web development technique for creating interactive Web applications. It is not a single new technology but combines a set of powerful, widely-used, well-known and mature technologies, mostly hosted by the client browser and largely independent of the server. Using AJAX technology greatly reduces the pressure on the server, improves the system response speed and reduces the waiting time. This paper describes its innovative mechanisms and elemental technologies in detail.**

***Index Terms*—Ajax, Asynchronous interaction, Seamless user application experience, GIS Crisis, Google Maps, Ajax Mobile Video**

I. Introduction

AJAX technology brings an increased responsiveness and interactivity of web pages by exchanging small amounts of data between the server and the web browser in the background without interfering with the display and behavior of the existing page. This achieves that entire web pages need not to be reloaded each time there is a need to fetch data from the server. XML HttpRequest, JavaScript and DOM are the core of AJAX technology.

Mobile technologies are rapidly changing our lives with increasing numbers of services supported by mobile devices, including Web-based learning applications, providing opportunities for people to study anytime and anywhere.

Using Web-based mobile applications to present learning resources is a challenge for developers because the performance of the mobile Internet over GPRS networks is often unacceptably slow. A new Web development model, Ajax, may help to address this problem. Ajax is an approach to Web application development that uses client-side scripting to reduce traffic between client and server and provide a seamless user application experience.

AJAX uses asynchronous interaction; AJAX introduces an intermediary between the user and server, then eliminates the shortcoming of processing - waiting - processing - waiting" in the network interaction, and improves the speed of data exchange of WEB application, updates on demand the contents of the application's user interface, responses rapidly of user behaviour of the browser client.

In the reminder of this paper, we summarize an innovative Ajax-like video technology. An asynchronous thumbnail data pre-fetching technique and a seamless media assembly and synchronization are introduced. Also we address the question of whether mobile Ajax provides measurable performance advantages over non-Ajax mobile learning applications. Here we discussed about the GIS Crisis-management systems using Ajax technology and Integrating Google Maps with GIS Visualization Systems, Also describes about AJAX technology applications in the network test system.

II. Ajax Technology Applications in the Network Test System

In the traditional WEB application, we interact by using the synchronization process; User action triggers a connection to the WEB server of HTTP requests, Server completes some processing, and then accesses other database systems, Finally, an HTML page returned to the client, when server is processing the request, the client user most of the time is in a wait state, When server has a heavy load, the server's response time is very long, this is a bad user experience. The AJAX (Asynchronous Javascript and XML) technology was born around 1999 with the introduction of the XMLHTTP object in Microsoft Outlook Web Access 2000 for accessing Web resources in background. The adoption of this concept in other systems, such as Mozilla browser with the support of the XML HTTPRequest object, opened the way to AJAX applications. The simple availability of a way to retrieve Web resources in background, interacting with them through a specific code, heavily modified the approach to Web systems design. Using AJAX, the client does not need to access information all at once since the required data can be retrieved in background if and when it is really needed. For example browsing a resource tree, the client can retrieve only the level that must be visualized. This substantially reduces the Web latency improving the overall system responsiveness.

A. Ajax Concept

AJAX is not a new technology, in fact it is several technologies that have been very mature in their respective fields, Includes the XML HttpRequest object, JavaScript, XML, XLST, DOM, XHTML, CSS, etc. Each technique has its unique, these technologies get together, become developed RIA (rich Internet application) mainstream technology. In these techniques: Standardizes presentation using XHTML

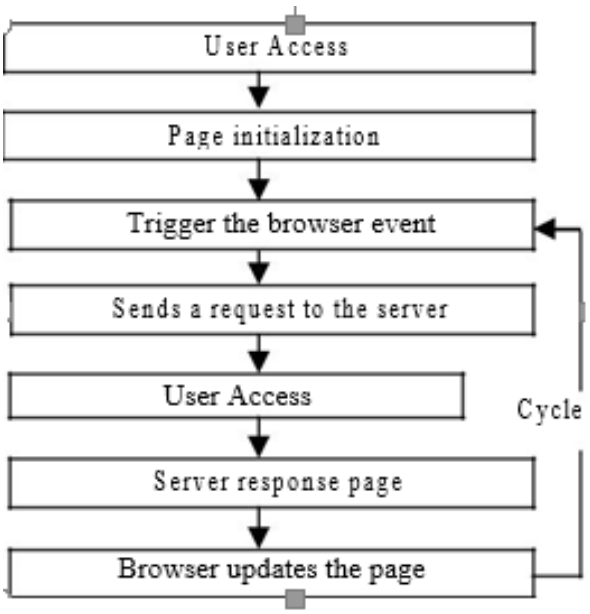

Fig. 1. Ajax life cycle

and CSS; Uses DOM for dynamic display and interaction; Uses XML and XSLT for data exchange and processing; Uses the XML HttpRequest object for asynchronous data reading; Finally, JavaScript binds the above techniques together.

III. Mobile Video Ajax Technology

The main core technology is a mobile video Ajax technology which can increase a light-weight, smooth, and quick responsiveness and interactivity of mobile video contents. This system pre-fetches and temporally stores discontinuous partial video portions in the future in the background of a users viewing behaviour. It allows viewers to jump to and re-start any desired scenes very quickly. In addition, it also provides a continuous and seamless video provisioning by switching between network streamed media and pre-fetched media. This combined access schemes is automatically synchronized based on viewers dynamic view context.

Scenarios of video viewing in a mobile environment have different features from those in conventional TV viewing or on-demand video viewing in static PC environments. This type of viewing has no strong intentions to view specific contents or programs in advance. Viewers tend to skip it frequently whenever they dislike it. Pre-downloading of video contents as files into mobile devices might be more preferable for fast video access. But the capacity of mobile device storage is quite low.it is essential to develop a new mechanism which can allow viewers to access desired contents smoothly and quickly in a short time, to skip or change contents easily, and to view a whole of content if they want without consuming device storages.

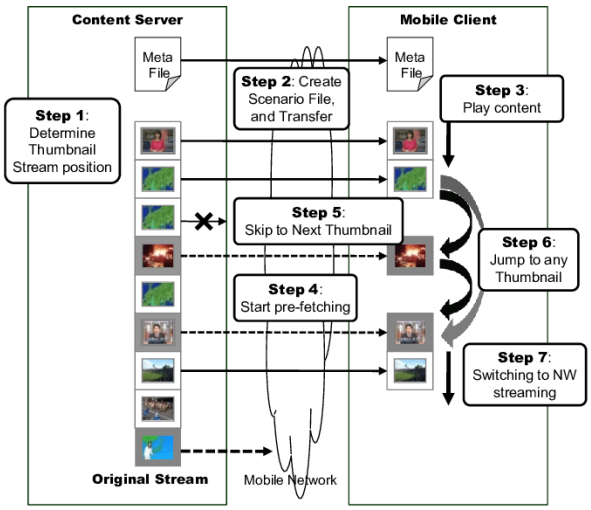

Fig. 2. Basic Mechanism

A. Basic mechanism

Light-weighted and fast content access capabilities are a must for mobile video services, where the user wants to jump very quickly to any desired scenes. However, it could be difficult to provide stable functionalities since mobile network conditions can be dynamically changed, and its network delay could be very high. This might impose a big burden on viewers skipping and jumping.The system adopts an asynchronous pre-fetching technology and seamless media assembly and synchronization technology. The following descriptions show how the proposed system works step by step as shown in Fig.2

1) The multi-positions of thumbnail stream inside one video stream.
2) This multi-positional information is stored in a metafile, and this file can be transferred into a mobile device prior to a real content playing phase.
3) viewer starts content playing on his/her mobile device by using network streaming mechanism.
4) Once viewing phase starts, pre-fetch procedure of thumbnail streams starts after a while. During this phase, the mobile client pre-fetches multi positional thumbnail streams by considering users viewing context.
5) quick and light-weight skipping and jumping phase. The mechanism allows to just skip to the top of the next portion of the prefetched thumbnail stream.
6) If the viewer wants to search other desired scenes, he/she can select a Selectable Jump Mode.
7) if the viewer continues to watch the stream, a successive stream after this thumbnail will be streamed via network.

B. View Context Aware Control Technologies

1) Asynchronous Thumbnail Prefetching: The proposed system pre-fetches accessible thumbnail streams asynchronously in the background of receiving streaming media that viewer is watching. This prefetching procedure follows

the meta file which describes the location and play timing of candidate thumbnail streams to be pre-fetched. This pre-fetching considering mobile network characteristics can retrieve thumbnail streams without disturbing streaming session. This additional functionality can give light-weighted and fast content access capabilities in a mobile environment without being interfered by unstable mobile network condition.

2) Seamless Media Assembly and Synchronization: It is required to assemble seamlessly two independent media sources, network streamed media and pre-fetched media files, without disturbing users viewing. As for media synchronization, the point is that network streamed media and thumbnail stream media can have completely independent timestamps. A unified timestamp is newly assigned into each media prior to media decoding process.

IV. Ajax Mobile Learning Systems

Ajax was originally introduced as a desktop Web programming model, not a mobile environment solution, but because of its key characteristics (small transmission volumes, asynchronous communication and partial Web page updates), it can actually be very useful in the mobile environment, especially in high latency networks where Ajax can reduce the frequency and volume of data transfer.

The traditional Web application allows users to submit a request, which the server will process and respond to. Then the client browser will refresh the whole Web page, even if only a small part of the content is changed. This can waste both network resources and users time. A traditional Web application architecture has two layers; user interface and server while Ajax has three layers; user interface, Ajax engine, and server.

The Ajax engine plays the role of middle-ware sitting between the user interface and the server. It controls the user interface and the data trans-mission between client and server, and also some logical functions. When the user initiates his or her request, this request will pass to the Ajax engine first, instead of the server directly as it does in a traditional client-server model. Then the Ajax engine will combine the data from the Web browser cache and data requested from the server to update the page. The main advantage of this combining data process is that it can reduce the size of data downloaded from the server, and increase the response speed. Ajax uses a browser hosted Ajax engine to handle both data transmission and partial updates to the Web page.

- Only requests new content from the server, reducing unnecessary data transmission.
- Partially update the current Web page when the response is received.
- Ajax engine runs locally in the client browser, the Web application's response speed is faster and the users experience is improved.
- In addition, client server communication in Ajax can be carried out asynchronously, enabling the user to continue interacting with the system even while the browser is waiting for data from the server.

The partial update and asynchronous features provide a number of advantages:

- Instead of submitting form data a page at a time when explicitly requested by the user, Ajax does submissions automatically, when the user triggers some event.
- Requests can be sent asynchronously, with the browser receiving results continually.
- Smaller, incremental in-place partial Web page updates can be made, instead of full page rebuilds.
- The size of data transmission can be reduced as the Ajax engine only downloads new content, not new pages.

A. Ajax v/s Non -Ajax

The studies show that there are differences in the performance between the Ajax and the non-Ajax mobile learning applications, users were asked to complete the quiz using the mobile device, Task one required users to finish all the multi-choice test questions in the Ajax mobile learning system, and task two required users to finish a similar set of multi-choice test questions using the ASP mobile learning system. After the evaluation of these systems in terms of their performance and usability, our results indicate that an Ajax mobile learning system can reduce data transmission volumes and the servers response time, and is preferred by users. An Ajax approach to Web-based mobile applications is a much more practical way to improve system performance than updating the mobile hardware or the wireless network.

Advantages:

- It allows users to access to the system as long as their mobile browser supports Ajax and has connectivity to a GSM/GPRS network.
- System performance is much faster than a traditional Web application.
- Reduced data traffic can save money for mobile learners in territories where mobile telecommunications companies charge for the amount of data sent and received.
- Requires little infrastructure dependence, because it is a software technology and does not require any hardware.
- Reducing the bandwidth required, and speeding up the user interface on the mobile device provides the user with a better mobile Internet experience.

V. GIS Crisis-Management Systems

The purpose of this system is to support restoration of disaster-affected facilities. This system has a GIS-Server that unies management of data from several maps, facilities data, digital road map (DRM) data, disaster information, and so on. These map data are raster data made from vector maps, and they were divided into tiled images by the rasterizing process on the server side. These raster maps are produced in every scale and managed by the server, which transfers images that are necessary for display through the Internet. And these maps are used as a background map. This Web-GIS displays maps to patch tiled raster images with the JavaScript program on the client side. If the map data is a vector, a user can continually zoom in/out. But this Web-GIS uses only raster

maps of previously dened scales, so a user cannot continually zoom in/out. And facilities information sent from the GIS-Server by asynchronous communication 2 is displayed on the map. Using Ajax technology, a user can repeat an operation without getting a reply from the server.

A. *Fast Map Response*

Ajax is a group of inter-related web development techniques used for creating interactive web applications. A primary characteristic is the increased responsiveness and interactivity of web pages so that entire web pages do not have to be reloaded each time there is a need to fetch data from the server. When a user operates a map, for example scrolling the map, conventional Web-GIS cant accept the users next operation. But if Web-GIS applies Ajax technology, a user can repeat an operation without getting a reply from the server. Fig.3 shows the difference between Classic Web-GIS and Ajax Web-GIS. In this way, this system offers fast-loading, asynchronous display updates, and smooths map scrolling. And the execution environment on the client side requires only a web browser (e.g. Internet Explorer, Firefox, Safari) without particular plug-in software (e.g. Java Runtime Environment, Flash), making this system highly convenient.

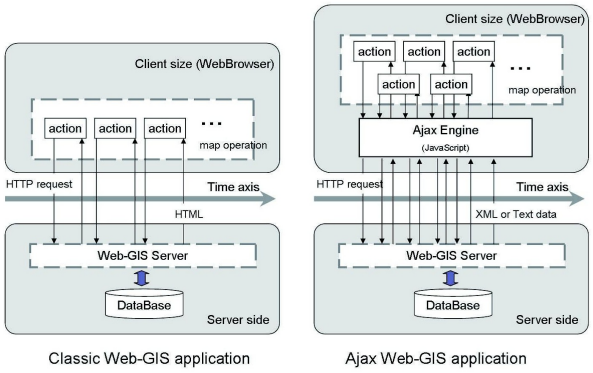

Fig. 3. Difference between Classic Web-GIS and Ajax Web-GIS.

VI. Integrating Google Maps with GIS Visualization Systems

Integration is basically coupling AJAX actions with the Web Services invocations, and synchronizing the actions and returned objects from the point of end users.

A. *Googles Ajax Integration with WMS*

There are two different paths working in parallel by the given user parameters created by the client actions. Actions are interpreted by the browser through the Google Mapping tools. JavaScript captures these actions by Action Listeners and Google Binding APIs and gives to Layer-2 object. On the browser user interface class is a JSP page. It includes two JavaScript class-references. One is for the Google Map object and the other is for the WMS map image and bindings to the Google Map object.

Interconnection for creating Layer-2 is done in accordance with the proposed architecture defined in Fig. 4. For Layer-1, a classic Google mapping application is used through the AJAX web application module and XML Http Request protocol. Google handles creating the map by using XML Http Request and given remote JavaScript file in the browser.

When we use this type of interaction interface to WMS, we can utilize all the OGC compatible functionalities of the WMS such as getMap, getCapabilities and getFeatureInfo. The client is going to be a thin client; it just takes the map and overlays it over the Google map. Overlay is done by using some advanced JavaScript techniques. The client does not need to make rendering or mapping jobs to create the map image. The map is already returned by the WMS and in a ready to use format such as JPEG or PNG or TIFF. Return type is defined as a parameter in the getMap request given to WMS. These images in different formats are converted to a JavaScript object before overlaying.

B. *Googles Ajax Integration with WFS*

WFS provides feature data in vector format and vector data are encoded in GML according to OGC WFS specifications and depending on the parameters given in the getFeature request. GML is an XML encoding for the transport and storage of geographic information, including both the geometry and properties of geographic features.

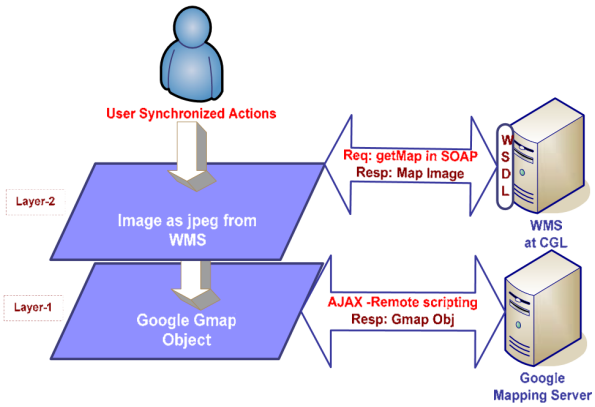

Fig. 4. Integration of Google Maps with OGC WMS.

VII. Conclusion

From the above studies have shown that Ajax reduces the data transmission volumes between the server and client device, and improves the user experience. A mobile video Ajax-like technology for light-weight and quick-responsible accessibility to contents was provided. Ajax mobile learning system can provide better performance in the mobile environment on the user experience level than non-Ajax mobile learning systems. Ajax Web-GIS offers fast-loading, asynchronous display updates, and smooths map scrolling.

REFERENCES

[1] Hokyoung Ryu and David Parsons , *Innovative mobile learning : techniques and technologies*

[2] Hiroyuki KASAI and Naofumi UCHIHARA ,*Mobile Video AJAX Technology for Time-Directional Quick Access*

[3] Xiaofeng Wang , *AJAX technology applications in the network test system*

[4] Adeolu Olabode Afoladi *"On Mobile Cloud Computing in a Mobile Learning System"*,Ladoke Akintola University of Technology,May 2014

[5] Goth, G., & Costlow, T. (2007). *"Currents - The Google Web Toolkit Shines a Light on"* Ajax Frameworks. Software, IEEE, 24(2), 94-98.

[6] GSM Association. (2003). *"GPRS Roaming Guidelines."* Retrieved August 29th, 2006, from http://www.gsmworld.com/technology/gprs/index.shtrnl

[7] GSM Association. (2007). *"What is GSM?"* Retrieved 6th August, 2007, from http://www.gsmworld.com/technology/what.shtml

[8] Sayar, A., Pierce, M., & Fox, G. (2006).*"Integrating AJAX Approach into GIS Visualization Web Services."*Paper presented at the Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/ Advanced International Conference.

Firewall: Design, Verification, Packet Filtering and Firewall Log Analysis

Aiswarya M R, Prashob Sasidharan and Vishnu Chandran C
Vidya Academy of Science & Technology
Thrissur - 680501, India

Manesh D
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: manesh.d@vidyaacademy.ac.in)

***Abstract*—A firewall is a device installed between the internet network of an organization and the rest of Internet. When a computer is connected to Internet, it can create many problems for corporate companies. Most companies put a large amount of confidential information online. Such an information should not be disclosed to the unauthorized persons. Second problem is that the virus, worms and other digital pests can breach the security and can destroy the valuable data.**

***Index Terms*—Firewall, packet filtering, log analysis**

I. Introduction

In computing a firewall is a network security system that monitors and controls incoming and outgoing network traffic based on predetermined security rules. A firewall typically establishes a barrier between a trusted internal network and untrusted external network, such as the Internet. The strength of the firewall is depends on the security level of the policies or rules in it. Sometimes the rules will overlap and made some confusion in the decision (accept or deny) making.

The use of Internet is increasing day by day in the modern era. The wide development of Internet leads to the cyber attack. Firewall is the most adequate security measure taken against the cyber attack. The effectiveness of firewall is measured by the performance that it gives. The two most important firewall products are packet filtering and application firewalls. Application firewalls inspect the contents of network packets while packet filter firewalls just inspect the head of each packet. Thus, packet filter firewalls usually have better performance than application firewalls .In this paper we deal with the packet filtering firewalls. Packet filtering firewall works in the Network layer.

In firewall, security policy is implemented based on the rules defined by the network administrator; that decides which packets can be allowed to an organizations private network. Manual definition of rules often results in anomalies in the policy. Therefore, an effective anomaly detection and resolution approach is needed. After resolving these conflicts, the rules can be re-ordered dynamically that improves the efficiency of the anomaly management framework. With firewall log analysis, frequently used rules can be set as primitive rules, to which more security can be added.

II. Designing

Here we discus about a new method of firewall designing. In this design the firewall consists of two parts, Single domain decision firewall rule and Binary tree firewall rule. These rules will helps to reduce the conflict in decision making of firewall.

In single domain decision, there are two types: a Close Firewall System (CFS), and an Open Firewall System (OFS). Firstly, CFS is an implicit denying for all services at the beginning of the start-up system. This system leads to all packets cannot pass through the firewall. The packets are always dropped. After that, an administrator can allow some necessary services.OFS always opens for all services to pass through the firewall at the start-up time. After working for a while, there may be some harmful activities, such as viruses, worms, denial of service, and so on. The administrator should check, and then block these services or activities.
The Binary Tree firewall is a data structure and an algorithm to fast check the firewall rules. It also supports the concept of SDD and it increases the speed from O(N2) of the general firewall to O(logN).

A. Current Problems of Firewall Rule Management

The most serious problem of firewall rule management is anomaly such as shadowing, correlation, generalization and redundancy. It occurs from several reasons. For instances, IT executives have not a good plan for IT management. Some companies with complex IT activities may generate inconsistent rules. These rules may come from carelessly adding some exception rules for some urgent activities, misunderstanding or an ignorance of organizational network policies.

B. Solution

Single Domain Decision concept (SDD) to solve rule anomalies, and the data structure for firewall rule (Binary Tree Firewall: BTF) to increase the speed of rule verification.

C. A Principle of SDD and Designing

In the basic of single domain decision, there are two types: a Close Firewall System (CFS), and an Open Firewall System (OFS). Firstly, CFS is an implicit denying for all services at the beginning of the start-up system. This system leads to all packets cannot pass through the firewall. The packets are always dropped. After that, an administrator can allow some necessary services. This system is quite reasonable for safety system. In the other hand, OFS always opens for all services to pass through the firewall at the start-up time. After working for a while, there may be some harmful activities, such as viruses, worms, denial of service, and so on. The administrator should check, and then block these services or activities.

III. Verification

Firewalls are the mainstay of enterprise security and the most widely adopted technology for protecting private networks. The quality of protection provided by a firewall directly depends on the quality of its policy (i.e., configuration). Due to the lack of tools for verifying firewall policies, most firewalls on the Internet have been plagued with policy errors. A firewall policy error either creates security holes that will allow malicious traffic to sneak into a private network or blocks legitimate traffic and disrupts normal business processes, which in turn could lead to irreparable, if not tragic, consequences.

Firewall verification tool takes as input a firewall policy and a given property, then outputs whether the policy satisfies the property. Despite of the importance of verifying firewall policies, this problem has not been explored in previous work. Due to the complex nature of firewall policies, designing algorithms for such a verification tool is challenging. we designed and implemented a verification algorithm using decision diagrams, and tested it on both real-life firewall policies and synthetic firewall policies of large sizes. A firewall is placed at the point of entry between a private network and the outside Internet such that all incoming and outgoing packets have to pass through it. The function of a firewall is to examine every incoming or outgoing packet and decide whether to accept or discard it. This function is specified by a sequence (i.e., an ordered list) of rules, which is called the policy, i.e., the configuration, of the firewall. Each rule in a firewall policy is of the form predicate to decision. The predicate of a rule is a Boolean expression over some packet fields such as source IP address, destination IP address, source port number, destination port number, and protocol type. The decision of a rule can be accept, discard, or a combination of these decisions with other options such as a logging option. The rules in a firewall policy often conflict. To resolve such conflicts, the decision for each packet is the decision of the first (i.e., highest priority) rule that the packet matches.

A. Firewall Verification Steps

1) Property Representation:
   To verify whether a firewall satisfies a given property, we need to translate the property to a set of non-overlapping rules, which we call property rules in this context to distinguish them from firewall rules.
2) Verification of Non-overlapping Firewall:
   To make our firewall verification algorithm easy to understand, we first assume that the given firewalls are non-overlapping. A firewall is non-overlapping if and only if no rules in the firewall overlap. Two rules overlap if and only if there exists at least one packet that can match both rules Note that real-life firewalls are most likely overlapping ones. The above unrealistic assumption is only for the purpose of introducing our verification algorithm that does not require this assumption.
3) Verification of Generic Firewalls:
   Here we consider the verification of generic firewalls, where the given firewall can be either overlapping or non-overlapping. A firewall is called an overlapping firewall if and only if there are at least two rules in the firewall which are overlapping.

$$
\begin{array}{lllllll}
r_1: & F_1 \in [20, & 50] & \wedge F_2 \in [20, & 70] & \rightarrow accept \\
r_2: & F_1 \in [1, & 60] & \wedge F_2 \in [40, & 100] & \rightarrow discard \\
r_3: & F_1 \in [1, & 100] & \wedge F_2 \in [1, & 100] & \rightarrow accept
\end{array}
$$

Fig. 1. An overlapping firewall

$$
\begin{array}{lllllll}
R_1: & F_1 \in [20, & 50] & \wedge F_2 \in [20, & 70] & \rightarrow accept \\
R_2: & F_1 \in [20, & 50] & \wedge F_2 \in [1, & 19] & \rightarrow accept \\
R_3: & F_1 \in [20, & 50] & \wedge F_2 \in [71, & 100] & \rightarrow discard \\
R_4: & F_1 \in [1, & 19] & \wedge F_2 \in [1, & 39] & \rightarrow accept \\
R_5: & F_1 \in [51, & 60] & \wedge F_2 \in [1, & 39] & \rightarrow accept \\
R_6: & F_1 \in [1, & 19] & \wedge F_2 \in [40, & 100] & \rightarrow discard \\
R_7: & F_1 \in [51, & 60] & \wedge F_2 \in [1, & 100] & \rightarrow discard \\
R_8: & F_1 \in [61, & 100] & \wedge F_2 \in [1, & 100] & \rightarrow discard
\end{array}
$$

Fig. 2. A non-overlapping firewall

B. Theorem (Firewall Verification Theorem)

The theorem states the following: "*A firewall decision diagram satisfies a property rule if and only if the property rule does not conflict with any rule defined by a decision path of the firewall decision diagram.*"

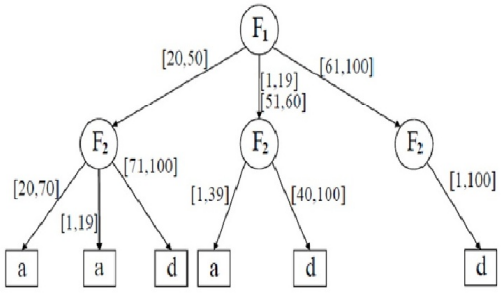

Fig. 3. A firewall decision diagram

Firewall Verification Algorithm
Input : (1)A firewall
(2)A property rule $(F_1 \in S_1) \wedge \cdots \wedge (F_d \in S_d) \rightarrow \langle dec \rangle$
Output : *true* if the firewall satisfies the property rule;
false otherwise.
Steps:
1. Convert the given firewall to a firewall decision diagram;
2. **return**(Verify($root, (F_1 \in S_1) \wedge \cdots \wedge (F_d \in S_d) \rightarrow \langle dec \rangle$));

Verify(v, $(F_1 \in S_1) \wedge \cdots \wedge (F_d \in S_d) \rightarrow \langle dec \rangle$)
1. **if** (v is a terminal node) and ($F(v) = \langle dec \rangle$)
then return true;
if (v is a terminal node) and ($F(v) \neq \langle dec \rangle$)
then return false;
2. **if** (v is a nonterminal node) **then**
/*Let F_i be the label of v*/
for each edge e in $E(v)$ **do**
if ($I(e) \cap S_i \neq \emptyset \wedge$
$\sim Verify(e.t, (F_1 \in S_1) \wedge \cdots \wedge (F_d \in S_d) \rightarrow \langle dec \rangle$))
then return false;
3. **return true**;

Fig. 4. Firewall verification algorithm

IV. Improving Performance of a Packet Filtering Firewall

The effectiveness of firewall is measured by the performance that it gives. The two most important firewall products are packet filtering and application firewalls. Application firewalls inspect the contents of network packets while packet filter firewalls just inspect the head of each packet. Thus, packet filter firewalls usually have better performance than application firewalls .In this paper we deal with the packet filtering firewalls. Packet filtering firewall works in the Network layer. In this type of firewall the focus is mainly on the individual packets .A packet filtering firewall is less secured compared with the other firewall. The problem is that they trust the packet themselves telling the truth who they are from and where they are from. The another main problem of packet filtering firewall is that it does not consider the traffic behaviour of the packets. Fuzzy Logic technique is used to overcome these problems and improve the performance of the firewall.

A. Fuzzy Perty Net

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. This section describes the kinds of mining activities that have been applied to the Web domain.

B. Methodology

1) Level 1 Fuzzy Filtering: This level depend on capturing and classifying all arrival packets depending on the information related with each packet such as IP address, packet time and protocol type to trace the packet movement. In this approach the packet is represented by a token in FPN. Once the packet is captured by a gate; it moves to place where checking and matching with ACL is done, in addition an instant copy of this packet is moved to the traffic analysis part to extract the packet's parameters (features) like number of packets of ICMP protocol coming through time period. These two parameters (packets counting and time period) embody the inputs to fuzzy logic engine that is used to produce the risk level. This risk level represents threats coming through movement of packets from untrusted sources.

The proposed system deals with ICMP protocol because ICMP has been used in many phases of an attackers advance in a system compromise. ICMP flood attacks (with characteristics similar to that of the Acknowledge flood) are traffic-based attacks that use heavy traffic to bring high loads to the servers, which will affect the server's normal services.

Once the system acquires the fuzzy descriptions of the packets' features, the rule base (fuzzy reasoning) can be built to make an inference of their similarity. Fuzzy reasoning, which is formulated by group of fuzzy IFTHEN rules, presents a degree of presence or absence of association or interaction between the elements of two or more sets. In the proposed system, reasoning is carried out through the following rules:

- Rule 1: If ICMP-echo-rate = high and time duration =long, Then Risk = moderate.
- Rule 2: If ICMP-echo-rate= medium and time duration= long, then Risk = moderate.
- Rule 3: If ICMP-echo-rate = low and time duration =long, then Risk =small.
- Rule 4: If ICMP-echo-rate = high and time duration =short, then Risk = large.
- Rule 5: If ICMP-echo-rate = medium and time duration= short, then Risk = large.
- Rule 6: If ICMP-echo-rate = low and time duration =short, then Risk = moderate.

The six rules altogether deal with the weight assignments impliedly in the same way as what humans think. The fuzzy inference processes all of the six cases in a parallel manner, which makes the decision more reasonable.

2) Level 2 Fuzzy Filtering: Typically, there are two sets of packets that associated with each firewall: the set of packets that are accepted by the firewall, and the set of packets that are discarded by the firewall . Our system invests this fact to enhance packet filtering performance by adopting level -2 fuzzy filtering to monitor rate of packets' acceptance or rejection to reorder. ACL rules with the aim of minimizing the time of rule matching . Here, the attempt is to model the uncertainty of the rate of acceptance or rejection of packets through a fuzzy model. In this case, a two input single-output fuzzy system is used. Two fuzzy variables including low, and high are used to describe both counter of acceptance rate Ar and rejection rate Rr. All of membership function' parameters are numerically specified based on the experiences and experimental results to adjust ACL rules ordering. In the same manner, the proposed system reasoning is carried out

through the following rules:

- Rule 1: If Ar = high and Rr = low, then Cr = high accept.
- Rule 2: If Ar = high and Rr = high, then Cr = equal,
- Rule 3: If Ar = low and Rr = low, then Cr = equal.
- Rule 4 If Ar = low and Rr = high, then Cr = high reject.

The outputs of fuzzy values are then defuzzified to generate a crisp value for the variable.

V. Firewall Log Analysis and Dynamic Rule Re-ordering in Firewall Policy Anomaly Management Framework

3) Firewall Policy Anomalies:

- Generalization:
  A rule is a generalization of one or more of preceding rules if they have different actions and if a subset of packets matched by this rule also matches the preceding rules.
- Shadowing:
  A rule is said to be shadowed when, one or more of preceding rules that matches all the packets matched by this rule, in such a way that the shadowed rule is never activated. Shadowing can be considered as a critical error in the policy, because the shadowed rule never takes effect
- Correlation:
  If a rule intersects with other rules but have different action, then this rule is said to be correlated with other rules. Here, the packets matched by the intersection of those rules may be denied by one rule, but permitted by others.
- Redundancy:
  A rule is redundant if there is another same or more general rule available that has same action on the same packet such that if the redundant rule is the overall firewall policy will not be affected.

A. System Achitecture and Design

A simple framework for anomaly detection and resolution is designed. The whole framework can be divided into two: an administrator end and a user end. System architecture is shown in figure. The administrator has to authenticate before performing any actions. After login, administrator have options for rule generation, anomaly detection and resolution and, rule re-ordering which is done dynamically analyzing the firewall log. The end user part selects the destination IP and destination port to which it has to send a file. And that is checked at the rule engine, with the defined firewall policy whether the transaction can be allowed or denied. Based on the user's transaction, firewall log is generated and the most frequently used firewall rules are mined that can be set as primitive rules.

B. Firewall Log Analysis

Firewall log analysis would generate a set of primitive rules with repeated and rare outcomes, which can be used to add more security for frequent rule in firewall log. Purpose

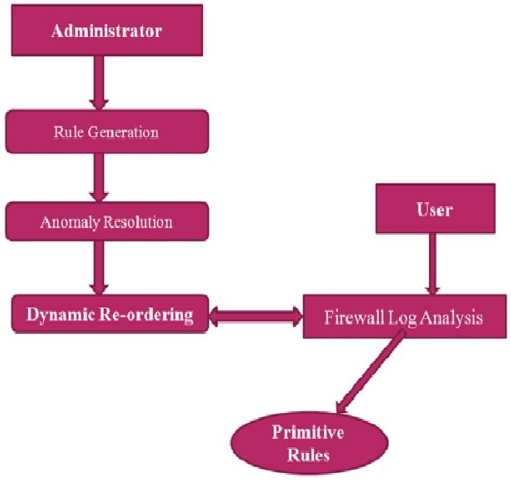

Fig. 5. Firewall log analysis

is to mine the firewall log data belonging to dynamical database accumulating as firewall system in operation. For e.g.:-Particularly, for corporate large network, firewall system would generates millions of log data every day, and further, the need of dealing with large amount of logs certainly imposes additional work burden on policy management system.

C. Psuedo Random Test Generator

1) Why Network Monitoring is Important: Today, most businesses rely on a computer and network infrastructure for internet, internal management, telephone and email. A complex set of servers and network equipment is required to ensure that business data flows seamlessly between employees, offices, and customers. The economical success of an organization is tightly connected with the flow of data. The computer network's reliability, speed, and efficiency are crucial for businesses to be successful. But, like all other technical objects, network devices may fail from time to time potentially causing trouble and loss of sales, no matter what migration efforts have been made up-front.

2) Monitor Networks with PRTG Network Monitor: PRTG Network Monitor is a powerful network monitoring application for Windows-based systems. It is suitable for small, medium, and large networks and capable of LAN, WAN, WLAN and VPN monitoring. You can also monitor real or virtual web, mail, and file servers, Linux systems, Windows clients, routers, and many more. It monitors network availability and bandwidth usage as well as various other network parameters such as quality of service, memory load and CPU usages. It provides system administrators with live readings and periodical usage trends to optimize the efficiency, layout and setup of leased lines, routers, firewalls, servers and other network components. The software is easy to set up and use and monitors a network using Simple Network Management Protocol (SNMP), Windows Management Instrumentation (WMI), packet sniffer.

Cisco NetFlow (as well as sFlow and jFlow) and many other industry standard protocols. It runs on a Windows-based machine in your network for 24-hours every day. PRTG Network Monitor constantly records the network usage parameters and the availability of network systems. The recorded data is stored in an internal database for later analysis

3) What PRTG Can Be Used For:

- Monitor and alert for uptimes/downtimes or slow servers.
- Monitor and account bandwidth and network device usage.
- Monitor system usage (CPU loads, free memory, free disk space etc.).
- Classify network traffic by source/destination and content.
- Discover unusual, suspicious or malicious activity with devices or users.
- Measure QoS and VoIP parameters and control service level agreements (SLA).
- Discover and assess network devices.
- Monitor fail-safe using a failover cluster setup.

4) Firewall Monitoring and Network Safeguarding with PRTG: The stability of the network firewall is responsible for the protection of both the processes which are running as well as sensitive data. It is of the greatest importance to corporate security that the firewall be stable. IT service interruptions can paralyze the company, while the theft of company figures, future plans, and customer data can have unforeseeable consequences. It is all the more vital, therefore, that the firewall remains uninterrupted and always up to date. With PRTG Firewall monitoring, you get a series of sensors which makes monitoring your firewall straightforward and effortless. Enjoy a clear overview of inbound and outbound network traffic, and be informed immediately in the event of a crash. This will enable you to promptly take action to restore the security of your system.

In addition, PRTG constantly monitors for firewall updates - for outdated firewall software is synonymous with a threat to security.

VI. Conclusion

This paper presents a method for formally verifying firewall policies. Such a tool is extremely useful in many ways. For example, it can be used in firewall debugging and troubleshooting. It also can be used iteratively in the process of designing a firewall. Firewalls are a very important component of system security. Recently, the packet filtering optimization problem has received the attention of the research community for many years. Firewall policy anomaly management is a complex task because of the large number of interacting rules in a firewall and is also error prone since rule generation and updation are done manually

References

[1] C. Tankard, “Advanced persistent threats and how to monitor and deter them,” Network Security, vol. 2011, no. 8, pp. 16- 19,2011.

[2] K. Ingham and S. Forrest, “A history and survey of network firewalls,” University of New Mexico, Tech. Rep, 2002.

[3] A. Liu, “Firewall policy change-impact ACM Trans. Internet Techno/., vol. 11, pp. 15:1-15:24, Mar. 2008. [Online]. http://doi.acm.org/IO.1145/2109211.2109212 analysis, no. 4, Available:

[4] G. Stoneburner, A. Goguen, and A. Feringa, “Risk management guide for information technology systems,” Nist special publication, vol. 800, no. 30, pp. 800-30, 2002.

[5] D. Hoffman, D. Prabhakar, and P. Strooper, “Testing iptables,”in Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research. IBM Press, 2003, pp. 80-91

[6] M. Ihde and W. H. Sanders, “Barbarians in the gate: An experimental validation of nic-based distributed firewall performance and flood tolerance,” in Dependable Systems and Networks, 2006. DSN 2006. International Conference on. IEEE, 2006, pp. 209-216.

[7] C. Kostnick and M. Mancuso, “Firewall performance analysis report,” Computer Sciences Corporation-SSC-NSD, Hanover MD,1995.

[8] H. Hu, G. J. Ahn, K. Kulkarni, “Detecting and Resolving Firewall Policy Anomalies, IEEE Transactions on Dependable and Secure Computing, 2012.

[9] H.C. van Assen, M.G. Danilouchkine, F. Behloul, H.J. Lamb, R.J.vanderGeest, J.H.C. Reiber, and B.P.F. Lelieveldt,Cardiac LV Segmentation Using a 3D Active Shape Model Driven by Fuzzy Inference,MICCAI 2003

[10] Passive Measurement and Analysis Project, National Laboratory for Applied Network Research. Auckland-VIII Traces. http://pma.nlanr.net/Special/ auck8.html”, December 2003.

[11] E. Al-Shaer and H. Hamed. “Design and Implementation of Firewall Policy Advisor Tools, School of Computer Science Telecommunications and Information Systems, Aug 2002.

Challenges and Security Issues in E-commerce and Solutions to Those Issues

Ajay Shankar, Giya Joy and Jereena K Francis
Vidya Academy of Science & Technology
Thrissur- 680501, India

Manesh D
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: manesh.d@vidyaacademy.ac.in)

***Abstract*—By the upcoming of technologies there has always been some security breaches in them and the same is the condition of e-commerce where security and privacy are two important elements. These two issues i.e. security and privacy are required to be looked into through social, organizational, technical and economic perspectives. In the C2C ecommerce the antifraud ability appears to be weak in general even though there exists reputation models to encourage good services of seller as well as punish their bad ones. Here in this study, by the application of our TRUST model and the method of fraud pattern recognition with time, we propose a fraud identification method. We also consider the challenges faced by the customer that they face due to lack of knowledge about the online transaction.**

***Index Terms*—E-commerce security, threats ,e-commerce cycle, transaction phases.**

I. Introduction

E-COMMERCE is one of the present technologies that is trending throughout the world. As every technology has its pros and cons e-commerce has its own privacy and security issues and various challenges that is faced by the users and the business peoples. In the present scenario the challenges faced by the consumers are very much relevant. Though there are many users world wide 70 percent of them are not aware of the risks and challenges that are hidden in the online sites. In considering how people might protect themselves from such risks, we should not underestimate the level of technology awareness that an end user really needs in order to be a safe e-Consumer. Although the largest volume and value of transactions occur in the business-to-business(B2B) context, it can be argued that the most challenging security issues are encountered in the scenarios that involve consumers. Whereas organizations engaging in B2B transactions can reasonably be expected to possess some level of competent IT support (and potentially even in-house security expertise), this is far less likely to be guaranteed when dealing with the end-user community in a business-to-consumer(B2C) scenario. As such, there are two potentially problematic issues for users when placed in an e-commerce.

- Scenario:
  - the need to recognise the threats.
  - the need to use the technology securely.

With the growing popularity of online trading sites reputation system are increasingly becoming an integral part of C2C ecommerce system. Reputation system can collect, aggregate and distribute participant feedback from past action to encourage sellers honest behaviours and effectively avoid cheating behaviours of those dishonest sellers.

The world largest C2C online author site eBay has a reputation system dealing with feedback information upon the completion of each transaction buyers and sellers have rights to give rating points (7,0,1)of the other. Each participant will have an identification name, and its evaluation will be given in connection with the transaction name on it. There reputation rating mechanisms cant well deal with the reputation slander ,the reputation speculation and other means of fraud generally.

In order to deal with the fraud patterns mentioned above, Based on TRUST model-¿A combined Reputation model for Distributed system, we propose a new fraud pattern identification and filtering method. It is to find fraud pattern in time window scope and filter out those fraud rating .

II. Transaction Phases

In e-commerce every transaction will have to go through many phases for a successful transaction to happen. Each phase also ensures different security measures that has to be implemented so that the transactions are not lost or fail.

III. Dimensions of E-commerce Security

There are different dimension of E-Commerce security which are thus:

1) Integrity: It is related with checking against unauthorized data modification and its reuse without users' permission and tempering of information.
2) Non-repudiation: Not to deny a sale or purchase
3) Authentication: It ensures that you are the one who is only allowed to logon to your IBA (Internet Banking Account)

4) Confidentiality: It is associated with encryption as well as decryption.
5) Privacy: The ability to control the terms under which personal information is acquired and used.
6) Availability: It is concerned with checking of data removal.
7) Auditing: Required to keep a record of operation.

IV. Common Mistakes Committed by Users

Novice errors committed by the users while engaged in online transaction help the fraudsters to take its fullest advantages. These common mistakes are:

1) Shopping on insecure websites
2) Divulging additional personal information which is not needed
3) Leaving computer open to virus

V. Two Common Patterns of Fraud

A. The Reputation Slander

It is a method in which ,some seller tries to degrade other seller by encouraging them in partnership or they register number of buyers to deliberately give low ratings to their competitors. In order to achieve we aim of suppressing their competitor.

B. The Reputation Speculation

In this method certain sellers and their accomplices conduct high score for each other or sellers register a great quantity of accounts performing virtual transaction to get high ratings.

VI. Anti-Fraud Methods

A. Trust Model

The basis of TRUST is the rating records from its buyers. Thus, each record of associated with sellers has a certain influence on the final reputation computation. The reputation value from TRUST is made of three steps, ratings collecting, ratings classifying and ratings computing. Weight is the most important factor in TRUST.

B. The Anti-Fraud Method

After indirect ratings are collected and classified, the ratings that lie in the time window are kept, others are thrown away. Then the fraud recognition arithmetic based on fraud pattern is run. Fraud ratings filtering program is in action after the fraud pattern analysis of ratings. Finally we compute the reputation value of the seller using TRUST.

C. Simulation Experiments

In order to verify the validity of TRUST + Anti-Fraud method, we implement a multi-agent system based on JADE that simulates the relationships and interactions between sellers and buyers in which reputation model help buyers to select sellers. Simulations are run in rounds. In the simulation environments, we achieve three kinds of reputation models: our TRUST+ Anti-Fraud method a typical centralized model CEN and another typical distributed model FIRE.

D. Involving the Reputation Slander

We set up 200 Slander agents in the simulation environment, and make them give low ratings to some sellers with high reputation values in a certain probability. In the 15th round of simulation the slander agents are activated. TRUST suddenly drops when the agents are activated, and then restore. FIREs curve declines slowly, CEN model, generally keep no change. As time goes on, due to the identification and filtering mechanisms, after more than 20 rounds of adjustments, TRUST will be able to return to the best condition, and the gains is very stable. FIRE does not consider the fraud problem; therefore, it will not recover after the decline in the yield curve.CEN incur with sudden drops.

E. Involving the Reputation Speculation

We set up 200 reputation speculation agents, giving them hype function. The agents will hype the sellers with higher ratings from 15 rounds .In this case TRUST, FIRE and CEN models are all proved to be beneficial to buyers. It shows that all reputation models mentioned above can help buyers to select profitable sellers to transact. However, TRUST outperforms FIRE and CEN, whether in fraud environment. Speculation differs from slander. Speculation is to hype those bad sellers to give them high ratings which harms the reputation system more than the slander does and makes system users benefit less.

VII. Security Challenges in Consumer Oriented E-commerce

There are so many technologies are available to protect e-commerce transactions. But consumers still face significant challenges because they have the lack of awareness of threats and lack of knowledge in the technology. The need to ensure threat awareness is well illustrated by the problem of phishing-a threat that can specifically target e-commerce users.

A. Understanding the Threats

In this section we discuss about the potential threats to e-commerce. The main issue is the consumers lack of knowledge of how it will work and what services that will provide. Good examples are phishing and pharming attacks, which attempt to dupe the user into divulging sensitive information such as credit card numbers, bank account details, and passwords and so on. Phishing attempts are typically initiated via email, with the user being send a fake message.

Pharming attacks have the same intention as phishing, but are conducted in a more complex manner. Whereas phishing relies upon tricking the user via a faked message, pharming involves poisoning the Domain Name System, with an attacker hijacking the domain name for a legitimate site and redirecting traffic to their site instead. Whereas users can be trained to avoid phishing attempts once they are familiar with the signs of a faked message, pharming attacks are difficult for them to avoid directly.

Phishing messages do not only spoof the brand, but can also fake security-related assurances in order to increase the

chances of gaining the users confidence. In addition to the potential for compromising personal data, phishing also undermines email as a viable means for ecommerce operators to pursue direct communication with their existing customers. So, if they want to have the facility to send legitimate and trusted messages to specific customers, merchants are obliged to use other means, such as incorporating messaging functionality within their websites

Phishing messages do not only spoof the brand, but can also fake security related assurances in order to increase the chances of gaining the users confidence.

B. Understanding the Technology

Whether they fully understanding threats or not, many users still recognize a requirement for security in an e commerce scenario. It is important to consider what they actually understand about the security they need or the protection they are receiving. Here we identified a variety of problems, which may collectively impede or undermine the level of protection that can be achieved.

C. Consistency of Functionality

Whereas we have striven for consistency in OS and application environments (such that users can develop skills that become transferable between different activities), the eCommerce experience can still be incredibly variable. . From a security perspective this could mean inconsistencies in behavior such as

- The type of user authentication mechanism in use.
- Whether the session will timeout after a period of inactivity.
- Whether the site will log them out simply by closing the browser, or whether an explicit sign out activity is required.

These variations will place demands upon what users need to know and remember in order to make secure use of different sites, and as a consequence they may cause confusion or lead to potential vulnerabilities.

The recognition that users cannot always be trusted to use passwords correctly leads some online service providers to use alternative means of authentication particularly where the account has significant potential for abuse if compromised. Another alternative is to introduce two-factor approaches, in order to present an additional obstacle to would-be attackers. A typical approach is to supplement the password with a further level of authentication, and an example here is the use of graphical passcodes, with users being authenticated on the basis of recalling a sequence of personal images .In this system, users select three personal images at the time they register for the system, and subsequent login attempts then incorporate these are part of the authentication process.

The user is faced with issues regarding logging out, and they again need to know the behaviour that the system expects of them. In some cases, the system attempts to safeguard against users who forget to log out by terminating sessions after a period of inactivity (e.g. online banking sites are amongst the most likely candidates to do this). However, if this becomes the users default assumption, then they risk becoming complacent about the need to sign out properly.

Many users will assume that closing the browser is sufficient to terminate their session. However, depending upon the site being used and the users browser configuration, many sites will attempt to store cookies on the system and thus maintain the users identity the next time someone navigates to the site from the same machine.

D. Misunderstanding the Protection

Legitimate merchants are well aware that consumer trust is important to the success of their eCommerce services. As such, it is now common to try to boost confidence by obtaining endorsements from third party providers in relation to the sites compliance with recognized security and privacy standards. The clear advantage here is that consumers no longer need to make a personal judgment about the credibility of a given site, and instead only require confidence in the credibility of the certifying party.

The anecdotal experiences suggest that users are satisfied to be told that they are secure, and have little reason to doubt the situation or question the nature of the protection. This again leaves uneducated users with a greater potential to be tricked, by lulling them into a false sense of security

VIII. Browser configurations

The users may find that usability issues inhibit appropriate use of security features. From the consumer perspective, two key requirements need to be fulfilled:

- the security options must make sense
- the system has to provide meaningful security-related feedback to its user.

Although these requirements are clearly not exclusive to eCommerce sites, they certainly have the potential to cause complications for users in this context, and examining the operation of popular browsers quickly reveals that these points are not adequately addressed.

If the user does manage to enable the functionality they require, a second potential problem could be that this renders their system vulnerable to other threats in the process particularly if they forget to revert to a more secure setting after completing their task.

From the usability perspective, setting the security too high can also be an obstacle for the unwary consumer. Having discovered the existence of security options, the natural inclination for those concerned about the issue may be to set the protection as high as possible. However, this does not always work to their advantage indeed, in some situations, it can prevent a site from working at all.

IX. Conclusion

In summery security and privacy are still ongoing research problems. In the last few years, there have been some important and interesting findings the researchers have reveled which have tried to throw light on the way to overcome the problems

of these security and privacy issues which are jeopardizing the trust factor. In spite of so many emphasis to combat the threat of security and privacy issues in online transactions, the security experts and unscrupulous hackers are found to be engaged in unwanted cat-and-mouse game. It is expected economics and sociologically best analysis will be able to ultimately bring greater transparency and proficiency in the online process so that the users can get rid of these issues of threat and E-Commerce business would assume its easy flow without any hindrance. However in this paper in a very simple form the matter has been discussed with pros and cons and a simple and thorough guideline has been proposed for the benefit of the users so that user can use online transaction in a very safe and secured mode.

REFERENCES

[1] Rashad Yazdanifard, Noor Al-Huda Edres, "Security and Privacy issues as a Potential Risk for Further E commerce Development", International Conference on Information Communication and Management - IPCSIT Vol. 16, 2011.

[2] Raju Barskar, Anjkna Jayant Deen, "The Algorithm Analysis of E-Commerce Security Issues for Online Payment Transaction System in Banking Technology", (UCSIS) - Vol.8, No.1, April, 2010.

[3] V. Srikanth, "E-commerce Online Security And Trust Marks", UCET ISSN 0976-6375, Vol. 3, Issue 2, July-September, 2012.

[4] Amitai Etzioni, *The Limits of Privacy*, New York, Basic Books, 1999.

[5] Sheshadri Chatterjee, "Security and Privacy Issues in E-Commerce: A Proposed Guidelines to Mitigate the Risk", 2015 IEEE International Advance Computing Conference (IACC).

[6] S.M.Furnell, "Considering the Security Challenges in Consumer-Oriented eCommerce", Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005..

[7] Fusheng Jin, Zhendong Niu, Haiyang Lang, "A Pattern Based Anti-Fraud Method in C2C Ecommerce Environment", 2010 International Conference on E-Business and E-Government.

[8] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood, "The Value of Reputation on eBay", *J. Experimental Economics*, vol.2, pp. 79101, 2006.

[9] N. Sundaresan, "Online Trust and Reputation Systems", *Proceedings of the 8th ACM conference on Electronic commerce*, ACM Press, New York, pp. 366367, 2007.

[10] A.Colley, "Phishing scam most devious ever", em ZDNet Australia, 3 March 2004.

A Study on Web Data Mining Types and Research Issues

Anil Augustine Chalissery, Soyet K Y and Stibin Varghese
Vidya Academy of Science & Technology
Thrissur - 680501, India

Reji C Joy
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: reji.c.j@vidyaacademy.ac.in)

***Abstract*—Web Data Mining is an important area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web, It can be classified into three different types i.e. web content mining, web structure mining and web usages mining. The aim of this paper is to provide past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining and also outlines key future research directions. This paper also reports the comparisons and summary of various methods of web data mining with applications, which gives the overview of development in research and some important research issues. Future trends of web data mining research have also been considered. In this paper, we have studied the basic concepts of web mining, classification, processes and issues**

***Index Terms*—Web data mining, web usage mining, web content mining, web structure mining**

I. Introduction

WEB MINING is the application of data mining technique which is an unstructured or semi- structured data and it automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web. The significant web mining applications are website design, web search, search engines, information retrieval, network management, E- commerce, business and artificial intelligence, web market places and web communities. Online business breaks the barrier of time and space as compared to the physical office business. Big companies around the world are realizing that e-commerce is not just buying and selling over Internet, rather it improves the efficiency to compete with other giants in the market. This application includes the temporal issues for the users.

Web mining has three classifications namely, web content mining, web structure mining and web usage mining. Each classification is having its own algorithms and tools. Web content mining is nothing but the discovery of valuable information from web documents and these web documents may contain text, image, hyperlinks, metadata and structured records. It is used to look at the information by search engine or web spiders i.e. Google, Yahoo. It is the process of retrieving the useful information from the web content or web documents. Web structure mining is also a process of discovering structured information from the websites. The structure of a graph consists of web pages and hyperlinks where the web pages are considered as nodes and the hyperlinks are edges and these are connecting between related pages. Web usage mining is also called as web log mining. It reflects the users behavior which can catch the meaningful patterns from one or more web localities

Web mining process consists of four important steps, they are, resource finding, data selection and pre-processing, generalization and analysis . Resource finding is the process which is used to extract the data either from online or offline text resources. In data selection and pre- processing step, specific information from retrieved web sources are automatically selected and pre-processed. During generalization, data mining and machine learning techniques are used to discover general patterns from individual web sites as well as across multiple sites. Validation and interpretation of the mined patterns are done in analysis step. . Web mining is classified into three different categories, they are, web content mining, web structure mining and web usage mining.

II. Web Data Mining

A. Overview

In 1996 its Etzioni who first coined the term web mining. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. According to Oren Etzioni Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and service. Web mining decomposes into the following sub tasks:

- Resource Discovery:
  locating unfamiliar documents and services on the Web.
- Information Selection and Pre-Processing:
  Automatically extracting and pre-processing specific information from newly discovered Web resources.
- Generalization:
  Uncovering general patterns at individual Web sites and across multiple Sites.

- Analysis:
  Validation and interpretation of mined patterns.
- Visualization:
  Presenting the results of an interactive analysis in a visual, easy to understand fashion.

B. Web Mining Categories

Web mining categories depend on which kind of data to be mined that is mining for information or mining the web link structure or mining for user navigation patterns.

Mining for information focuses on the development of techniques for assisting a user in finding documents that meet a certain criterion that is web content mining.

Web content mining refers to the discovery of useful information from web contents, including text, image; audio, video, etc..mining the link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining.

Web structure mining tries to discover the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites.

Finally, mining for user navigation patterns focuses on techniques which study the user behavior when navigating the web that is web usages mining.

Web usage mining refers discovery of user access patterns from Web servers. Web usages data include data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls or any other data as result of interaction.

C. Advantages of Web Mining

Web mining can obviously be quite beneficial to both businesses and individuals . Web mining is attractive for companies because of several advantages. In the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with the customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles.

D. Research Issues in Web Mining

The web is highly dynamic; lots of pages are added, updated and removed everyday and it handles huge set of information hence there is an arrival of many number of problems or issues. Normally, web data is high dimensional, limited query interface, keyword oriented search and limited customization to individual users. Due to this, it is very difficult to find the relevant information from the web which may create new issues.

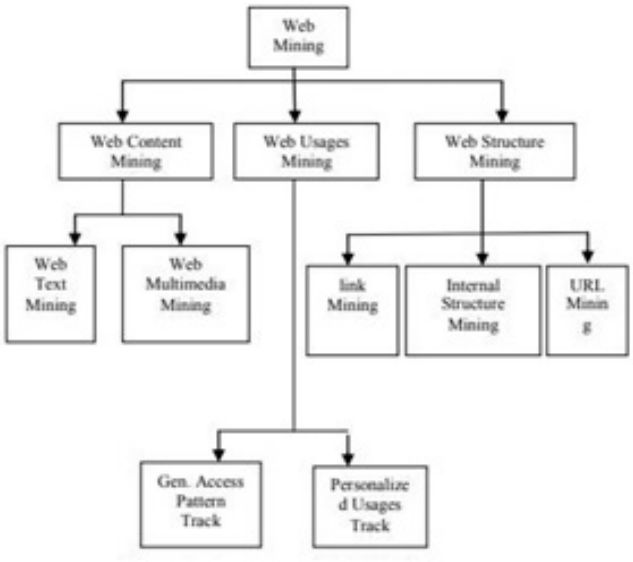

Fig. 1. Taxonamy of web mining

Web mining techniques are classification, clustering and association rules which are used to understand the customer behaviour, evaluate a particular website by using traditional data mining parameters. Web mining process is divided into four steps; they are resource finding, data selection and pre-processing, generalization and analysis.

Web measurement or web analytics are one of the significant challenges in web mining. The measurement factors are hits, page views, visits or user sessions and find the unique visitor regularly used to measure the user impact of various proposed changes. Large institutions and organizations archive usage data from the web sites. The main problem is that, detecting and/or preventing fraud activities. The web usage mining algorithms are more efficient and accurate. But there is a challenge that has to be taken into consideration. Web cleaning is the most important process but data cleaning becomes difficult when it comes to heterogeneous data . Maintaining accuracy in classifying the data needs to be concentrated. Although many classification techniques exist the quality of clustering is still a question to be answered.

E. Major Issues in Web Mining

Web data sets can be very large, it takes ten to hundreds of terabytes to store on the database.

- It cannot mine on a single server so it needs large number of server
- Proper organization of hardware and software to mine multi-terabyte data sets
- Limited customization, limited coverage, and limited query interface to individual users
- Automated data cleaning
- Over fitting and Under fitting of data

- Over sampling of data
- Scaling up for high dimensional data
- Mining sequence and time series data
- Difficulty in finding relevant information
- Extracting new knowledge from the web

III. Web Usage Mining

Usage Data is the data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses. Web Usage mining is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications.

A. Data Sources

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns.

1) Server Level Collection: A Web server log is an important soure for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log files can be stored in various formats. Cached page views are not recorded in as server log. In addition, any important information passed through the POST method will not be available in a server log. Packet sniffing technology is an alternative method to collecting usage data through server logs. Packet sniffers monitor network traffic coming to a web server and extract usage data directly from TCP/IP packets. The Web server an also store other kinds of usage information such as cookies and query data in separate logs. Such as Common log or Extended log formats.

2) Client Level Collection : Client-side data collection can be implemented by using a remote agent (such as JavaScript or Java applets) or by modifying the sour e ode of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser. Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems.

3) Proxy Level Collection : A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

B. Preprocessing

Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery. Usage preprocessing is arguably the most difficult task in the Web Usage Mining process due to the incompleteness of the available data. Unless a client side tracking mechanism is used, only the IP address, agent, and server side link stream are available to identify users and server sessions.

Assuming each user has now been identified (through cookies, logins, or IP/agent/path analysis), the link-stream for each user must be divided into sessions. Since page requests from other servers are not typically available, it is difficult to know when user has left a Website. A thirty minute time out is often used as the default method of breaking a user's link stream into sessions .When a session ID is embedded in each URI, the definition of a session is set by the content server. While the exact content served as a result of each user action is often available from the request field in the server logs, it is sometimes necessary to have access to the content server information as well. Since content servers can maintain state variables fo reach active session, the information necessary to determine exactly what content is served by a user request is not always available in the URI. The final problem encountered when preprocessing usage data is that of inferring cached page references .The only verifiable method of tracking a held page views is to monitor usage from the client side. The referrer held for each request can be used to detect some of the instances when cached pages have been viewed.

C. Pattern Discovery

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. This section describes the kinds of mining activities that have been applied to the Web domain.

1) Statistical Analysis: Statistical techniques are the most common method to extract knowledge about visitors to a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path.

2) Association Rules: Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks

3) Clustering: Clustering is a technique to group together a set of items having similar characteristics. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users.

4) Classification: Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular lass or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc.

5) Sequential Patterns: The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns includes trend analysis, change point detection, or similarity analysis

6) Dependency Modeling: Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain. . There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Such techniques include Hidden Markov Models and Bayesian Belief Networks. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption. Such information may help develop strategies to increase the sales of products offered by the Website or improve the navigational convenience of users.

D. Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, an often highlight overall patterns or trends in the data

E. Research issues in web usage mining

- Session identification
- CGI data
- Catching
- Dynamic pages
- Robot detection and filtering
- Transaction identification

IV. Web Content Mining

Web content mining though uses data mining techniques; it differs from data mining because Web data are mostly unstructured and/or semi-structured, while data mining deals mainly with structured data. It is associated to text mining because much of the Web contents are texts. Web content mining differs from text mining because of the semi structure quality of the Web, while text mining deals with unstructured texts. Web content mining thus requires inventive applications of text mining and/or data mining techniques and also its own distinct approaches.

A. Web Mining Tasks

1) Information Retrieval:
   It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web.
2) Pre-processing:
   It is the task of automatically selecting and pre-processing specific information from retrieved Web resources.
3) Pattern Recognition and Machine Learning:
   It is the task to automatically discover general patterns of individual Web sites as well as across multiple sites.
4) Analysis:
   It is the task of analyzing, validating and interpreting the mined patterns.

B. Web Content Mining Techniques

- It identifies the useful information from the web contents/data/documents.
- Web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents.
- Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

1) Unstructured Data Mining Techniques: Content mining has been accomplished on unstructured data such as text. Mining of unstructured data provides unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Some of the useful techniques used in text mining are as follows:

1) Information Extraction:
   The pattern matching technique is used to extract information from unstructured data. In this case, keyword and phrases are traces out and then connections with the keywords are found within the text. This technique is very useful when there is large volume of text. Information Extraction is the basis of many other techniques used for unstructured mining
2) Information Visualization:
   It utilizes feature extraction and key term indexing to build a graphical representation. The documents having similarity are determined using Information Visualization. Large textual materials are represented as visual hierarchy

or maps where browsing facility is allowed. It helps the user to visually analyze the contents.

3) Topic Tracking:
This technique checks the documents viewed by the user and studies the user profiles. According to each user it predicts the other documents related to users interest. In Topic Tracking applied by Yahoo, user can give a keyword and if anything related to the keyword pops up then it would be informed to the user. -The demerit of this technique is that when we search for topics we may be provided with information which is not related to our interest.
4) Summarization:
It has been used to reduce the length of the document by maintaining the main points. The time taken by the technique to summarize the document is less than the time taken by the user to read the first paragraph. To understand the important key points, summarization tool search for headings and sub headings to find out the important points of that document An example for text Summarization is Micro Soft Words Auto Summarize.
5) Categorization:
This technique is used to identify main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document.
6) Clustering:
This technique has been used to group similar documents. -Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering technique helps the user to easily select the topic of interest. Clustering technology has been used in Management Information Systems

2) Structured Data Mining Techniques:

1) Web Crawler:
Web Crawler such as: Internal and External Web Crawler. Internal Crawler crawls through internal pages of the Website. External Crawler crawls through unknown Website.
2) Page Content Mining:
Which works on the pages ranked by traditional search engines.
3) Wrapper Generation:
The wrappers will also provide a variety of Meta Information. I.e. domains, statistics, index look up about the sources.

3) Semi-Structured Data Mining Techniques:

1) Object Exchange Model:
Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange Model (OEM). A main feature of object exchange model is self describing; there is no need to describe in advance the structure of an object
2) Top down Extraction:
It extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted.
3) Web Data Extraction Language:
Web data extraction language converts web data to structured data and delivers to end users. It stores data in the form of tables

C. Multimedia Data Mining Techniques

1) SKICAT:
Image processing and data classification which helps to classify very large classification set. It uses machine learning technique to convert these objects to human usable classes.
2) Multimedia Miner:
Multimedia Miner contains four major steps, Image excavator ,preprocessor for extraction search kernel , discovery module
3) Colour Histogram Matching:
It contains Colour Histogram Equalization and Smoothing
4) Shot Boundary Detection:
It is a technique which automatically detects boundaries shots in the Videos

D. Study of Web Content Mining Tools

1) Automation Anywhere: Automation anywhere is a web data extraction tool. The Intelligent Automation Software, used for automating and scheduling business process and IT tasks in easier way. Features of Automation Anywhere

- Intelligent automation is used for business and IT tasks
- Distributes tasks to multiple computers easily, using Task to SMART Exe capability
- Web recorder: (Used for extracting multiple Data and to extract Table)

2) Web Info Extractor: This tool is helpful in mining web data, extracting web content, and monitoring content update. for content retrieval it is a very powerful tool. It can retrieve unstructured or structured data from web page, reorganize into local file or save to database, place into web server Difficult template rules are not required to be defined.

3) 3Web Content Extractor: It is the most powerful and easy-to-use data extraction tool for web scraping, data mining or data extraction from the internet. It tool allows users to extract data from various websites such as online stores, online auctions, shopping sites, real estate sites, financial sites, business directories, and etc. The extracted data can be exported to a variety of formats. including Microsoft Excel (CSV), Access, TXT, HTML XML, SQL script, MySQL script and to any ODBC data source.

4) Screen-Scraper: Screen-scraping is a tool for extracting information from web sites which can be used in other contexts. It allows mining the content from the web, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements.

5) Mozenda: To extract web data easily and to manage it affordably Mozenda is useful.

E. Research Issues on Web Content Mining

Web content mining has number of research issues because it can extract the information from the web search engines. Data / Information Extraction concentrate on extraction of structured data from web pages such as products and search results.

Web information integration and schema matching. The web contains large amount of data, each website accept similar information in a different way. Similar data discovery is an important problem with lots of realistic applications. Opinion extraction from online sources i.e. customer makes sure of products, forums, blogs and chat rooms. Mining opinions are of big consequence for marketing intelligence and product benchmarking.

Automatically segmenting web pages and detecting noise is an interesting problem in web application. It could not have advertisements, navigation links and copyrights notices. Hence, extracting the main content of the web page is important problem in web application.

V. Web Structure Mining

Web structure mining is the study of data interconnected to the structure of a particular website. It consists of web graph which contains the web pages or web documents as nodes and hyperlinks as edges those are connecting between two related pages .

Web structure is useful source for extracting information. Web structure is to extract some interesting web graph patterns like co-citation, social choice, complete bipartite graphs, etc. It classifies the web page on various topics and deciding which web page is to be added into the collection of web pages. Web structure mining can be performed either at intra-page level or inter-page level. A hyperlink that connects to a different part of the same page is called intra-page hyperlink. It is a document structure level.

A hyperlink that connects two different pages are called inter-page hyperlink which is structure level . Web page is organized in tree structure format based on HTML tags. Here, the documents are extracted automatically by the Document Object Model (DOM). The main reason for developing link mining is to understand the social organization of the web. The research of structure analysis is called Link mining which is located in the connection of work in link analysis, hypertext and web mining, relational learning, inductive logic programming and graph mining. Some of the important tasks of link mining are link based classification, link based cluster analysis, link type, link strength and link cardinality. The research of the hyperlink level is also called hyperlink analysis which can be used to retrieve useful information from the web.

Web structure mining is used in search engines such as Google, Yahoo, etc. HITS algorithm was used in clever search engine by IBM and the page rank algorithm is used by Google . Algorithms of web structure mining are HITS (Hypertext Induced Topic Search) algorithm, Max flow- Min cut algorithm, ECLAT algorithm, and Page rank algorithm. Page rank algorithm can be divided into two types. One is weighted page rank algorithm and another one is Topic sensitive page rank algorithm.

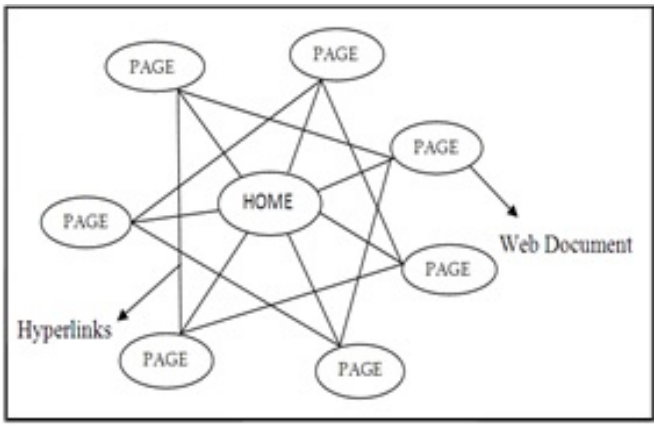

Fig. 2. Web graph structure

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining , which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet.

The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web Structure Mining.

1) Link-based Classification:
   The most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.
2) Link-based Cluster Analysis:
   The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

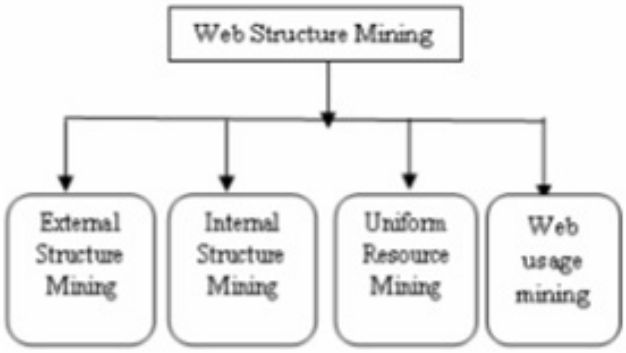

Fig. 3. The hierarchical structure for web structure mining

3) Link Type:
   There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.
4) Link Strength:
   Links could be associated with weights.
5) Link Cardinality:
   The main task here is to predict the number of links between objects, Page categorization, finding related pages, finding duplicated web sites and to find out similarity between them.

A. Functions of Web Structure Mining

Figure below represents the hierarchical structure for Web Structure Mining.

From the above Figure the Hierarchical structure for Web Structure Mining is also known as Link Analysis process. The research of structure analysis had increased in value and focuses on future research concept with scope and we named called as Link mining. The Web contains a variety of stuff with almost no identical or standard structure, and it differs in the design / style and as well the content to be better than in usual collections of text documents. The uses of web structure mining in among the online users.

1) It used to rank the online users queries.
2) Used to finding the related web pages from the website
3) Finding the similarity between the websites and its category.
4) Improving navigation of web pages on business websites.
5) It used to mine the previously unidentified link between web pages.
6) Discovering the structure of web document.
7) Structure mining can be used to reveal the structure (Schema) of web pages.

B. Research Issues on Web Structure Mining

Web structure mining has two issues due to its huge amount of data. Reducing irrelevant search results. Relevance of search information becomes unorganized due to the problem search engines often only tolerate for low precision criteria. Indexing information on the web. This causes low amount of recall with content mining.

VI. Conclusion

This paper has provided a more current evaluation and update of web mining research available. Extensive literature has been reviewed based on three types of web mining, namely web content mining, web usage mining, and web structure mining. . Web data mining is a fast rising research area today. As the web data and its usage will rise in future. It will prolong to generate more content, structure and usage data. So the importance of web data continues increasing. Web data is mainly semi-structured to unstructured. Due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research Problems. This paper also has discussed about the research issues and challenges in web mining and also provided detailed review about the basic concepts of web mining, web content mining, structure mining, usage mining, tools, algorithms and types. Several open research issues and drawbacks which are exists in the current techniques are also discussed. This study and review would be helpful for researchers those who are doing their research in the domain of web mining.

References

[1] Gaurav Kumar, Pradeep Kumar Bhatia, “A Detailed Review of Feature Extraction in Image Processing Systems”, 2014 Fourth International Conference on Advanced Computing & Communication Technologies.
[2] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, 2nd ed., Beijing: Publishing House of Electronics Industry, 2007.
[3] T. Shraddha, K. Krishna, B.K.Singh and R. P. Singh, “Image Segmentation: A Review”, International Journal of Computer Science and Management Research Vol. 1 Issue. 4 November 2012.
[4] M.R. Khokher, A. Ghafoor and A. M. Siddiqui, “Image segmentation using multilevel graph cuts and graph development using fuzzy rule-based system”, IET image processing, 2012.
[5] V. Dey, Y. Zhang and M. Zhong, “A review on image segmentation techniques with Remote sensing perspective”, ISPRS, Vienna, Austria, July 2010.
[6] S. Inderpal and K. Dinesh, “A Review on Different Image Segmentation Techniques”, IJAR, Vol. 4, April, 2014.
[7] Suzuki K., “False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs”, Academic Radiology, volume 13, February 2005.
[8] Gaurav Kumar, Pradeep Kumar Bhatia, “A Detailed Review of Feature Extraction in Image Processing Systems”, 2014 Fourth International Conference on Advanced Computing & Communication Technologies.
[9] H.C. van Assen, M.G. Danilouchkine *et al.*, “Cardiac LV Segmentation Using a 3D Active Shape Model Driven by Fuzzy Inference”, MICCAI 2003
[10] Olivier Ecabert, Jochen Peters *et al.*, “Automatic Model-Based Segmentation of the Heart in CT Images”, IEEE
[11] H.C. van Assen, M.G. Danilouchkine *et al.*, “Cardiac LVSegmentation Using a 3D Active Shape Model Driven by FuzzyInference”, MICCAI 2003.

E-Commerce Security Threats and Solutions

Anjali Anto, Poulin Davis V and Sijisha V S
Vidya Academy of Science & Technology
Thrissur - 680501, India

Dijesh P
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: dijesh.p@vidyaacademy.ac.in)

***Abstract*—E-commerce stands for "electronic commerce". E-commerce is a powerful tool for business transactions. Electronic commerce can help organization to reduce costs and to make greater market relationships between buyers and sellers. E-commerce Security is actually part of the Information Security framework and is specifically applied to the components that affect e-commerce that include Computer Security, Electronic commerce security has its own gradation and is one of the highest visible. This paper analyzes the threat classification and security measures,. Organizations using e-commerce can use the framework to improve their security. The purpose of this research paper is to explain the importance of E-commerce and to explain the security.**

***Index Terms*—E-Commerce, Security, Threats, Solutions**

I. Introduction

E-COMMERCE stands for Electronic Commerce. E-commerce is a powerful tool for business transactions. Electronic commerce can help organization to reduce costs and to make greater market relationships between buyers and sellers. E-Commerce is nothing but exchange of goods and services over the Internet. E-commerce Security is a part of the Information Security and is could be applied to the components that affect e-commerce that include Computer Security, E-commerce security is the protection of e-commerce assets from unauthorized access, use, alteration, or Destruction. It is necessary to build a secure system.

II. Advantages of E-Commerce

The advantages of e-commerce can be summarised as follows.

- Organizations can expand their market to worldwide markets with minimum investment.
- Customers can enquire about a product or service and place orders anytime, anywhere from any location.
- E-commerce helps the organizations to reduce the cost and manage the paper based information by digitizing the information.
- E-commerce helps organization to provide better customer services.
- E-commerce helps to simplify the business processes and makes them faster and efficient.
- E-commerce provides users with multiple options to compare and select the cheaper and better options.
- Customers need not travel to shop a particular product, thus there is less traffic on road and also less air pollution.
- E-commerce also helps in rural areas to access services, these services are not otherwise available to them.

III. Disadvantages of E-Commerce

E-commerce has some disadvantages also. The following are some of the major disadvantages.

- Lack of system security.
- Sometimes network bandwidth might cause an issue in some places.
- Sometimes, it becomes difficult to integrate an e-commerce software or website with existing applications or databases.
- There could be chances of software or hardware compatibility issues.
- Consumers may not trust the site, it is very difficult to convince user to use online mode to purchase.
- It is difficult to trust the security or privacy on online transactions.
- Lack of touch or feel of products during online shopping is one of the biggest drawbacks.
- Internet access is still not cheaper and is inconvenient to use for many customers, those are living in remote villages.

IV. E-Commerce Security Issues and Solutions

Security is essential in internet shopping. Nearly anything can be purchased, for example, music, toys, autos, nourishment and even dress. Despite the fact that some of these buys are illicit we will concentrate on all the things you can purchase legitimately on the web. A portion of the well known sites are eBay, iTunes, Amazon, HMV, Mercantile, dell, Best Buy and substantially more.

V. Purpose of Security

The purpose of security in E-shopping is that the payment transactions is safe or not, and it is highly require. Now a day, Most of the transactions are online based or internet based. We cannot say all the transactions are safely successful or not. Attacker may be attack the networks and stole the information. These are the basic concepts to understand the purpose of security.

1) Data confidentiality
2) Authentication and identification
3) Access Control
4) Data Integrity
5) Non-repudiation

VI. Security Issues

1) Authentication:
   Checks who you say you are. It implements that you are the just a single permitted to log on to your Internet saving money account.
2) Authorization:
   Permits just you to control your assets in particular ways. This keeps you from expanding and the adjust of your record or erasing a bill.
3) Encryption:
   Manages data covering up. It guarantees you can't keep an eye on others amid Internet during banking transactions
4) Auditing:
   Keeps a record of operations. Traders utilize examining to demonstrate that you purchased a particular stock.
5) Integrity:
   Prevention against unauthorized data modification.
6) Non-repudiation:
   Prevention against any one party from reneging on an agreement after the fact.
7) Availability:
   Prevention against data delays or removal.

VII. Security Threats in E-commerce

In E-commerce Threats may be occurs because attacker may be attack on the network during the transitions or payment time. There are some basics threats may be occur are as follows

1) Denial of Service (DOS Attack):
   There is no actual intent to cause damage to files or to the system, but the goal is to literally shut the server down. This happens when a massive amount of invalid data is sent to the server
2) Spamming:
   Sending spontaneous business messages to people E-mail bombarding brought about by a programmer focusing on one PC or system, and sending a huge number of email messages to it. Surfing includes programmers putting programming specialists onto an outsider framework and setting it off to send solicitations to a proposed target.
3) Viruses:
   A virus needs a host of some sort in order to cause damage to the system. The exact definition is a virus attaches itself to executable code and is executed when the software program begins to run or an infected file is opened.
4) Worms:
   Worms are very much different. A worm does not need a host to replicate. Rather, the worm replicates itself through the Internet, and can literally infect millions of computers on a global basis in just a matter of hours.
5) Trojan Horses :
   A Trojan Horse is a piece of programming code that is layered behind another program, and can perform covert, malicious functions. Programming and working frameworks' security gaps

VIII. E-Commerce Security Tools

There are various tools are used to provide the security.

1) Digital Signatures:
   We can use the digital signature for authenticate the person or user. Digital signatures are unique for each user.
2) Encryption software and techniques:
   It also use for security purpose because we want to pass the encrypted data on the network if the attacker attack the networks information will be secure into the form of chipper text there are various encryption software available in the present market. Example software: Bitlocker, veracrypt, Axcrypt.
3) Public key infrastructure:
   Public key infrastructure is a set of rules, policies and procedures needed to create manage, distribute use public key encryption.
4) Firewalls:
   A firewall is a network security system, either hardware- or software-based, that uses rules to control incoming and outgoing network traffic. (A firewall acts as a barrier between a trusted network and an un trusted network. A firewall controls access to the resources of a network through a positive control model. This means that the only traffic allowed onto the network is defined in the firewall policy; all other traffic is denied.)
5) Algorithm:
   Algorithms are also used for encrypted the data there are various algorithm to change the plaintext into chipper text. Algorithim:RC4, CRT.
6) Password schemes:
   Vital information can be protected by using passwords. It is widely used in network security. In password schemes, generally eight character length mixed case alphanumeric characters are chosen as password. The majority of hackers access client computers because of easy passwords.
7) Biometric systems:
   Biometric System is considered as the most secured of security methods. In this method, unique aspects of a persons body are taken as a recognition pattern. E.g. finger

prints, palm prints, retinal patterns of eyes, signatures or voice recognition.

8) Use of anti-virus software:
Client must always use the protection method and that is to scan for malicious data and program fragments that are transferred from the server to the client, and filter out data and programs known to be dangerous.

IX. Policies of E-Commerce Security Measure

There are different policies used to ensure and measure security in E-commerce environment, we shall explain some of them in the following sections, which are: Privacy, Cryptography, and certificates.

A. Privacy Policy

According to a study released by commerce Net & Nielsen Media Research, More than 2 out of every five people in North America are now Internet users, & the web is becoming as integral part of daily life. Without a through privacy security policy, its not possible to spend money in a responsible and cost effective manner. Develop a privacy security policy that includes defining the sensitivity of information, the exposure of the organization if that information was likelihood of those risks becoming reality. A policy may contain many elements including purchasing guidelines, statements of availability and Privacy.

B. Cryptography

Cipher systems are classified into two classes.

1) Secret key cipher system: Secret key cryptography is the oldest type of method in which to write things in secret. There are two main type of secrete key cryptography,*transposition* and *substitution. Transposition cipher* encrypts the original message by changing characters order in which they occurred. Where as in *substitution cipher,* the original message was encrypted by replacing the characters with other characters. In both types, both the sender and receiver share the same secret keys. The most widely used secret key scheme today is called Data Encryption Standard (DES). DES cipher work with 56-bit secret key and 16 rounds to transform a block of plaintext into cipher text.

2) Public Key cipher system: Public-key cryptography was developed to solve the secret-key distribution problem associated with secrete key method. It was first publicly described in 1976 by Stanford University Professor Martin Hellman and graduate student Whitfield Diffie. Public key method use two different keys. One of the keys is used to encrypt the data, i.e. plaintext and the second key is used to decrypt the cipher text .The second problem that , Diffie pondered, and one that was apparently unrelated to the first was that of digital signatures.

Rivest-Shamir-Adleman (RSA) scheme is the most widely accepted and implemented general-purpose approach to public-key encryption. The RSA scheme is a block cipher in which the plaintext and ciphertext are integers between 0 and $n-1$ for some n. A typical size of n is 1024 bits, or 309 decimal digits. The block size must be less than or equal to $\log_2(n)$. Encryption and decryption are of the following form, for some plaintext M, and cipher text C:

$$C = M^e (\text{mod } n)$$
$$M = C^d (\text{mod } n)$$

Both sender and receiver must know the value of n. The sender knows the value of e, and only the receiver knows the values of d. Thus, this is a public-key encryption algorithm with a public key of

$$\text{KU} = \{e, n\}$$

and a private key of

$$\text{KR} = \{d, n\}.$$

C. Certificate

Certificates bind identity, authority, public key, and the other information to a user. For most internet E-commerce application, certificates using a format defined in international telecommunication union telecommunication standardization sector (ITU-T). Recommendation X.509 is employed. An X.509 certificate contains such information as the following:

1) Certificate holders name and identifier.
2) Certificate holders public key information.
3) Key usage limitation definition.
4) Certificate policy information.
5) Certificate issuers name and identifier.
6) Certificate issuers name and identifier.

In todays E-commerce environment, buyers may get personal certificates to prove their identity to a web site but it is the vendor sites that really need to have certificates to prove their identity to buyers.

X. Secure E-Shopping Guidelines and Solution

There are several guidelines for secure E-shopping are as follows

1) Shop at secure web sites:
Firstly check whether the site is secure or not. Web website address is shown (the "address bar"), you ought to see https://. The "s" that is shown after "http" demonstrates that Web webpage is secure. Frequently, you don't see the "s" until you really move to the request page on the Web website.
2) Research the web site before you order:
In the event that you choose to purchase something from an organization, begin with a modest request to learn if the organization is dependable. Dependable organizations ought to publicize their physical place of work and no less than one telephone number, either client Benefit or a request line. Call the telephone number and make inquiries to decide whether the business is honest to goodness. Maybe companions or relatives who live in the city recorded can confirm the legitimacy of the organization. Recall that, anybody can make a Web webpage.

3) Read the web site's privacy and security policies:
   Each legitimate online Web webpage offers data about how it forms your request. It is typically recorded in the segment entitled Privacy Policy. You can see whether the vendor means to impart your data to an outsider or member organization. Do they require these organizations to abstain from advertising to their clients If not, you can hope to get spam (spontaneous email) and even mail or telephone requesting from these organizations. You can likewise realize what sort of data is assembled by the Web webpage, and how it is or is not imparted to others. The online dealer's information security practices are additionally frequently clarified in the Privacy Policy, or maybe a different Security Policy.
4) Whats safest: Credit cards, debit cards, cash, or check:
   The most secure approach to shop on the Internet is with a Visa. In the occasion something turns out badly, you are ensured under the government Fair Credit Billing Act. You have the privilege to question charges on your Visa, and you can withhold installments amid a lender examination. When it has been resolved that your credit was utilized without approval, you are in charge of the main $50 in charges.
5) Keep your password private:
   Numerous web based shopping destinations require the customer to sign in before putting in or seeing a request. The customer is normally required to give a secret word. Never uncover your secret word to anybody. While choosing a secret word, don't utilize usually referred to data, for example, your birth date, mother's last name by birth or numbers from your driver's permit or Social Security number. Try not to reuse a similar secret key for different locales, especially destinations related with delicate data. The best secret key has no less than eight characters and incorporates numbers and letters.

XI. Conclusion

E-commerce is growing at a tremendous speed with the excessive amount of data sharing and the tremendous number of users connected. There are enormous opportunities for unauthorized users to access confidential data. Security is an important factor in virtual transactions. It is the protection of any device, computer or any network from unauthorized access. This paper highlights the importance and basic needs of e-security authenticity, integrity, non repudiation, and authorization.

References

[1] Mazumdar Sengupta C. and Barik M.S., "E-commerce security-a life cycle approach", *Sadhana*, vol. 30, no. 2-3, (2005).
[2] Xiangsong M.and Fengwu H., "Design on PKI-based anonymous mobile agent security in e-commerce", *Wuhan University Journal of Natural Sciences*, vol. 11, no. 6, (2006).
[3] A. & R. Anindya "New threats to online banking", *PCQuest*, pp.62-67, July, 2005.
[4] Ackerman, Mark S., Lorrie Cranor, and Joseph Reagle, "Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences" in *Proceedings of the ACM Conference in Electronic Commerce*: pp.1-8.
[5] David J. Olkowski, Jr., "Information Security Issues in ECommerce", *SANS GIAC Security Essentials*, March 26,2001.
[6] Paul A. Greenberg, "In E-Commerce We Trust ... Not", *Ecommerce Time*, February 2, 2001, Available: `http: www.ecommercetimes.com/perl/story/?id=7194`.

Security Issues and Solutions in Cloud Computing

Anne Mariya Joseph, Haripriya V H and Sruthi P N
Vidya Academy of Science of Technology
Thrissur-680501, India

Sajay K R
Associate Professor of Computer Applications
Vidya Academy of Science of Technology
Thrissur-680501, India
(email: sajay.k.r@vidyaacademy.ac.in)

***Abstract*—Cloud computing has raised IT to newer limits by contribution the market environment data storage and volume with springy scalable computing processing power to match elastic demand supply, with reducing capital expenditure. Usually cloud computing services are delivered by a third party provider who owns the infrastructure. Cloud computing offers an innovative business model for organizations to adopt IT services without upfront investment. Security is one of the major issues which hamper the growth of cloud. Today, leading players, such as Amazon, Google, IBM, Microsoft and salesforce.com offer their cloud infrastructure for services. Hence, a requirement to re-consider security, privacy and trust considerations within the context of the cloud computing paradigm arises.**

***Index Terms*—OS, system, domain**

I. Introduction

Cloud computing is the fastest growing technology, offers various services over the internet. It can serve many facilities to the business such as resources, infrastructure, platform etc by paying amount on demand basis over network with the functionality of increase or decrease the requirements.

This technology can meet any IT requirements at any time. It can serves most of the hardware and software facilities required for companies for storing, creating, managing, running consumer applications on cloud in lease or rent basis, it provides resources as a service to multiple consumers by virtualization. This technology helps many IT organizations to start up business without huge economical barriers, slowly move to leading organization in the industry. It can serve facilities irrespective of the size of organizations. Information is stored in centralized servers and cached temporarily on clients that can include desktop computers, notebooks, handhelds, and other devices.

The complexity of cloud can be reduced by simply reducing it into replicated thousands of primitives and common functional units. These complexities create many issues related to security as well as all aspects of Cloud computing. Cloud typically has single security architecture but have many customers with different demands. The challenge of security issues arising due to the fact that both customer data and program are residing in Provider Premises.

The main objective of this paper is to give techniques to ensure cloud security.

II. Challenges of Cloud Computing

Cloud computing offers many benefits as mentioned above, even though cloud computing has many challenges. While moving from traditional computing to cloud computing, companies must aware about the benefits and challenges of cloud computing. While analyzing these challenges, security of data is the most tedious work in cloud computing. According to a survey carried out by Gartner, more than 70% of Chief Technical Officers believed that the primary reason for not using cloud computing services is that of the data security and privacy concerns. Convincing the organizations especially small ones about security concern is a tedious work; they are not ready to throw away their infrastructure and immediate move to cloud. Most of the organizations are closely watching this issue and not ready to shift to cloud space, this is the main reason in the lack of maturity level of cloud computing.

Some of the security challenges are discussed below.

1) Privacy of data
2) Confidentiality of data
3) Data remanence
4) Data integrity
5) Transmission of data
6) Malicious insiders

A. Privacy of Data

Privacy of data is key security concern for cloud computing. Most of organizations feeling more comfort while putting valuable data in their site than cloud space. Consumers do not have any idea regarding the location of data, transfer of date, operations on cloud, etc. Most of the organizations are unaware of security mechanism implemented by service providers. Many questions are arising by consumers such as the following:

1) Which are the organizations sharing services?

2) How creation and back-up of files taking place?
3) What happens to the deleted files?
4) Which type of consumers can access data?
5) Location of data?

B. Confidentiality of Data

Confidentiality is related to data privacy; it ensures data is visible to only authorized users. It is very difficult due to the virtualization and multi tenancy properties that multiple consumers sharing the hardware, software simultaneously in a distributed network. Confidentiality is the responsibility of service provider. Common solution to the confidentiality is encryption. Many symmetric and asymmetric algorithms are available for data confidentiality, even though encryption and decryption is the solution to the confidentiality, there are many questions are arising related to this.

1) Where is encryption and decryption taking place (client side or cloud side)?
2) How can search the data in an encrypted form?
3) What are threats while transferring data from client to cloud?
4) Any misuse of data by service provider?
5) Any misuse of key by service provider?

C. Data Remanence

Data should be deleted from cloud after the life-cycle, or the memory should be reformatted or recycled. The reformatting of storage media does not remove the previously written data from the media, but also it can be accessed or recovered from the media later. No clear standard is available for recycle the storage media. This data remanence makes difficult the vacation of hardware. Most consumers are unknown of allotted resources and storage space, due to this issue consumers are locked in one service provider. Various techniques have been developed to counter data remanence. These techniques are classified as cleaning, purging/sanitizing, or destruction. Specific methods include overwriting, degaussing, encryption, and media destruction

D. Data Integrity

Preservation of information from loss or modification by unauthorized users is referred as data integrity. Multiple organizations are sharing the application or platform by multi-tenancy, consumers working on same work may share data can be modified by any other unauthorized user sharing the application or platform in the cloud, this cause the integrity failure. As data are the base for providing cloud computing services, such as Data as a Service, Software as a Service, Platform as a Service, keeping data integrity is a fundamental task.

E. Transmission of Data

Most of the time data is transferring between consumer and cloud. Initially data is sent from client site to cloud, data is returned from cloud to client after queries during the operation. Encryption is used provide protection while the transmission of data. Most of the time data is transferred without encryption due to lot of time is required for encryption and decryption for each operation upon data. During transfer an attacker can trace the communication, interrupt the data transfer, miss use of data, etc. Homomorphic algorithm allows to process data in an encrypted form, even though there is a chance of data transfer interruption, change the data transfer, other issues.

F. Malicious Insiders

Malicious insiders are authorized employees, these users appointed for managing and maintaining cloud by cloud service provider. These users sometimes steal or corrupt the sensitive data of organizations in the cloud and convey this sensitive information to other organizations sharing the same cloud. These malicious insiders may get payment for this malicious work. Sometimes service provider not able to take any action against these employees.

III. Solutions to Security Issues

Some of the solutions for the security issues are:

1) Encryption:
   It ensures the confidentiality of the data stored in cloud
2) Authentication:
   It ensures the integrity of the Stored in cloud
3) Backup and recovery:
   It ensure the availability Of the data stored in the cloud

These are discussed in detail below.

IV. Encryption

It uses the concept of cryptography. Cryptography It is a science used to secure sensitive data. Confidentiality is the fundamental security service provided by cryptography, keeping data invisible to unauthorized users. Components of cryptosystem are follows:

1) Plaintext:
   Original form of data, data to be protected during transmission and storage.
2) Cipher text:
   It is the unreadable form of the plaintext after encryption operation.
3) Encryption algorithm:
   Used to convert plaintext to cipher text, it is a mathematical process.
4) Decryption algorithm:
   It performs reverse operation of encryption algorithm, convert cipher text to plaintext.
5) Encryption key:
   It is a value used by sender with algorithm to convert plaintext to cipher text.
6) Decryption key:
   It is a value used by receiver with algorithm to convert cipher text to plaintext.

Encryption algorithms have vital role in the field of cloud security. Many algorithms are available for cloud security. Most useful algorithms for cloud security are discussed below.

A. The Data Encryption Standard (DES)

The DES is a symmetric key block cipher published by the National Institute of Standards and Technology (NIST). It uses single key (secret key) for both encryption and decryption. It operates on 64-bit blocks of data with 56 bits key. The round key size is 48 bits.Entire plaintext is divided into blocks of 64bit size; last block is padded if necessary. Multiple permutations and substitutions are used throughout in order to increase the difficulty of performing a cryptanalysis on the cipher. DES algorithm consists of two permutations (P-boxes) and sixteen Feistel rounds. Entire operation can divided into three phase. First phase is Initial permutation and last phase is the final permutations.

1) Initial permutation rearranges the bits of 64-bit plaintext. It is not using any keys, working in a predefined form.
2) There are 16 fiestel rounds in second phase. Each round uses a different 48-bit round key applies to the plaintext bits to produce a 64-bit output, generated according to a predefined algorithm. The round-key generator generates sixteen 48-bit keys out of a 56-bit cipher key.
3) Finally last phase perform final permutation, reverse operation of initial permutation and the output is 64-bit cipher text.

B. The Advanced Encryption Standard (AES)

The AES is a symmetric key block cipher published by the National Institute of Standards and Technology (NIST). Most adopted symmetric encryption is AES. It operates computation on bytes rather than bits, treats 128 bits of plaintext block as 16 bytes. These 16 bytes are arranged in four columns and four rows for processing as a matrix. It operates on entire data block by using substitutions and permutations. The key size used for an AES cipher specifies the number of transformation rounds used in the encryption process. Possible keys and number of rounds are as following:

1) 10 rounds for 128-bit keys
2) 12 rounds for 192-bit keys.
3) 14 rounds for 256-bit keys.

Major advantages of AES over DES are the following:

1) Data block size is 128 bits.
2) Key size 128/192/256 bits depending on version.
3) Most CPUs now include hardware AES support making it very fast.
4) It uses substitution and permutations. 5. Possible keys are 2128 , 2192 and 2256.
5) More secure than DES.
6) Most adopted symmetric encryption algorithm.

C. Rivest-Shamir-Adleman (RSA) Scheme

The RSA scheme is a public key cipher developed by Ron Rivest, Adi Shamir and Len Adlemen in 1977. It is most popular asymmetric key cryptographic algorithm. This algorithm uses various data block size and various size keys. It has asymmetric keys for both encryption and decryption. It uses two prime numbers to generate the public and private keys. These two different keys are used for encryption and decryption purpose. This algorithm can be broadly classified in to three stages; key generation by using two prime numbers, encryption and decryption.

RSA today is used in hundreds of software products and can be used for key exchange, digital signatures, or encryption of small blocks of data. This algorithm is mainly used for secure communication and authentication upon an open communication channel.

When we use small values of p and q (prime numbers) selected for the designing of key, then the encryption process becomes too weak and one can be able to decrypt the data by using random probability theory and side channel attacks. On the other hand if large p and q lengths are selected then it consumes more time and the performance gets degraded in comparison with DES. Operation speed of RSA Encryption algorithms is slow compare to symmetric algorithms, moreover it is not secure than DES.

D. Homomorphic Algorithm

It is an encryption algorithm that provide remarkable computation facility over encrypted data(cipher text) and return encrypted result. This algorithm can solve many issues related to security and confidentiality issues. In this algorithm encryption and decryption taking place in client site and provider site operates upon encrypted data. This can solve threat while transferring data between client and service provider, it hide plaintext from service provider, provider operates upon ciphertext only.

Homomorphic encryption allows complex mathematical operations to be performed on encrypted data without using the original data. For plaintexts $X1$ and $X2$ and corresponding ciphertext $Y1$ and $Y2$, a homomorphic encryption scheme permits the computation of $X1 \circ X2$ from $Y1$ and $Y2$. The cryptosystem is multiplicative or additive homomorphic depending upon the operation "$\circ$" which can be multiplication or addition.

E. Blowfish Encryption Algorithm

Blowfish is a symmetric encryption algorithm designed in 1993 by Bruce Schneier as an alternative to existing encryption algorithms. Blowfish has a 64-bit block size and a variable key length from 32 bits to 448 bits. It is a 16-round Feistel cipher and uses large key-dependent S-boxes. While doing key scheduling, it generates large pseudo-random lookup tables by doing several encryptions. The tables depend on the user supplied key in a very complex way. This approach has been proven to be highly resistant against many attacks such as differential and linear cryptanalysis. Unfortunately, this also means that it is not the algorithm of choice for environments where a large memory space is not available. Blowfish is similar in structure to CAST-128, which uses fixed S-boxes.

Since then Blowfish has been analyzed considerably, and is gaining acceptance as a strong encryption algorithm. Blowfish was designed in 1993 by Bruce Schneier as a fast, free alternative to existing encryption algorithms. Since then it has been

analyzed considerably, and it is slowly gaining acceptance as a strong encryption algorithm. Blowfish is unpatented and license-free, and is available free for all uses. The only known attacks against Blowfish are based on its weak key classes.

F. Elliptic Curve Cryptography (ECC)

Elliptic Curve Cryptography (ECC) provides similar functionality to RSA. Elliptic Curve Cryptography (ECC) is being implemented in smaller devices like cell phones. It requires less computing power compared with RSA. ECC encryption systems are based on the idea of using points on a curve to define the public/private key pair.

G. El Gamal

El Gamal is an algorithm used for transmitting digital signatures and key exchanges. The method is based on calculating logarithms. El Gamal algorithm is based on the characteristics of logarithmic numbers and calculations. The Digital Signature Algorithm (DSA) is based on El Gamal algorithm.

1) Firebase Authentication: This type of authentication provides backend services, app SDKs, and libraries to authenticate users to a mobile or web app. This method authenticates users, using a variety of credentials like Google, Facebook, Twitter or GitHub. The Firebase authentication method uses a client library to sign a JSON Web Token, JWT, with a private key after the user has successfully signed in. This method then validates the JWT, through a proxy, was signed by Firebase and that the issuer matches the setting in API configuration.

2) AuthO Authentication: This method not only authenticates and authorizes apps and APIs but it is also stack, device, and identity agnostic. This method supports several providers and security assertion markup language specification. Much like Firebase Authentication, this method also provides backend services, SDKs and user interface libraries for authenticating users in web and mobile apps. Also, like Firebase Authentication, this method validates the JWT was signed and the issuer matches the API configuration.

3) Google Authentication: This authentication method allows users to authenticate by signing in with their Google account. Once the user is authenticated, they have access to all Google services and a Google ID token can be used to make calls to Google APIs and Cloud Endpoints APIs. This method also verifies that the JWT was signed by Google and the issuer is listed on the API configuration.

4) Google Authorization and Service Accounts: With this method, a JWT can be generated and signed using a service account and Google-provided client library for a Google Cloud Platform project. This method uses the public key to validate a Google-signed JWT and to ensure that Google is listed as the issuer in the API configuration. For this method, Google ID tokens are recommended for service accounts because the API producer only needs to whitelist Google as an issuer for all service accounts.

H. Cloud Computing Authentication Issues

1) Privacy issues
2) Lack of transparency
3) Security issues
4) The possibility of exploitation of the authentication mechanism
5) Different authentication technologies presents challenges to customers

When it comes to cloud computing, service providers require customers to store their account information in the cloud, giving service providers access to this information. For many customers, this presents a privacy issue for them. The lack of transparency in the cloud makes it difficult for customers to ensure the proper rules are enforced. Customers using multiple cloud services have more copies of their information out there in the cloud. This causes security issues for customers and cloud service providers. Multiple copies of accounts lead to multiple authentication processes and provide the possibility to exploit the authentication mechanism. Cloud service providers use different authentication technologies for authenticating users and while this has less of an impact on SaaS than PaaS and IaaS, it presents challenges to customers.

Benefit of Cloud Authentication Services is Cloud Authentication Services keep personal information secure. Many people and companies keep confidential information on their cloud servers. With all this personal information in a non-physical area, it is beneficial to have an authentication service to ensure this information stays classified.

There are many benefits of cloud authentication services including speed and flexibility. There are many cloud authentication service platforms connecting groups, devices, applications, and networks together.

I. Benefits of Cloud Authentication Services

1) Integration with RSA authentication manager
2) Methods to protect authentication manager resources
3) Convenient authentication
4) Centralized access control policies
5) Enforced security requirements for applications
6) Multifactor authentication support
7) Connectivity to popular applications
8) Single Point Access to Protected Applications
9) On-premise server user passwords
10) Passwords not synchronized to the cloud authentication service
11) Multifactor, multistep authentication
12) Interface definition able to be integrated with any programming language
13) Run-time authentication for protected resources
14) Able to modify and improve authentication capabilities
15) Runs on platform hosted through a global network
16) Segregates customer data to ensure privacy
17) Database environment shares infrastructure
18) Standardized service levels and operational procedures for all customers

When it comes to using the cloud for private information, many companies are concerned others will be able to access their information. Cloud Authentication Services help put

minds at ease with its ability to protect this information. The deployment of a Cloud Authentication Service consists of four main components: the on-premise Identity Router, the Cloud Administrative Console, the RSA SecurID Authenticate app installed on user devices, and the Cloud Authentication Service, which is the name of the managed cloud server and the set of components.

V. Cloud Backup

Cloud backup is commonly termed ass online backup. Cloud backup is a process of backing up your data to a cloud based remote server, often temed as clod storage. Google dire, Dropbox, Right Backup etc are common examples of cloud backup.

A. Cloud Backup Types

Backups may be categorized as belonging to one of the following types:

1) Full backup
2) Incremental backup
3) Differential backup
4) Mirror backup
5) Disk backup

1) Full Backup: As the name implies the full backup is when every single file and folder in the system is backed up.a full backup takes longer and require more space than other types of backups but the process of restoring lost data from backup is much faster

2) Incremental Backup: With incremental backup, only the initial backup is a full one. Subsequent backups only stores changes that were made since the previous backup. The process of restoring lost data from backup is longer but the backup process is much quicker.

3) Differential Backup: Differential backup is similar to incremental backup. With both, the initial backup is full and subsequent backups only store changes made to files since the last backup. This type of backup requires more storage space than incremental backup does, however, but it also allows for a faster restore time.

4) Mirror Backup: A mirror backup, as the name implies, is when an exact copy is made of the source data. The advantage of mirror backup as opposed to full, incremental, or differential backups, is that youre not storing old, obsolete files. When obsolete files are deleted, they disappear from the mirror backup as well when the system backs up. The downside to mirror backup is that if files are accidentally deleted, they can be lost from the backup is well if the deletion isnt discovered before the next scheduled backup

VI. Authentication

There are multiple authentication techniques in cloud computing suited for different applications and use cases when it comes to the cloud. The best cloud authentication method depends on your preferences but each is a supported method.

A. Cloud Authentication Methods

1) API Keys: This method doesnt require client libraries and is transparent to the user. This method identifies the project by creating a strong association between a key and a project. API keys are less secure as they are vulnerable to man-in-the-middle attacks. API keys can easily be added to any HTTP call as a query parameter in the header because they dont require a client library.

VII. Conclusion

Cloud computing is emerging as a new thing and many of the organizations are moving toward the cloud but lacking due to security reasons. So cloud security is must which will break the hindrance the acceptance of the cloud by the organizations. There are a lot of security algorithms which may be implemented to the cloud. DES, Triple-DES, AES, and Blowfish etc are some symmetric algorithm. DES and AES are mostly used symmetric algorithms.DES is quite simple to implement then AES. RSA and Diffie-Hellman Key Exchange is the asymmetric algorithms. In cloud computing both RSA and Diffie Hellman Key Exchange is used to generate encryption keys for symmetric algorithms. But the security algorithms which allow operations (like searching) on decrypted data are required for cloud computing, which will maintain the confidentiality of the data.

Authentication ensures the integrity of the data, that is, it prevents the manipulation and updation of the data by unauthorised persons, and backup and recovery ensures the availability of the data at anytime from anywhere and it protect the data from data loss server failure or from natural disasters.

References

[1] Mohiuddin Ahmed, Abu Sina Md. Raju Chowdhury , Mustaq Ahmed, Md. Mahmudul Hasan Rafee, "An Advanced Survey on Cloud Computing and State-of-the-art Research Issues, *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, No 1, January 2012 ISSN (Online): 1694-0814.

[2] Jason, "Defining Cloud Deployment Models, `http://bizcloudnetwork.com/defining-cloud-deploymentmodels`, Last modified on August 4, 2010.

[3] Ang Li, Xiaowei Yang, Srikanth Kandula and Ming Zhang, "Comparing Public Cloud Providers, *IEEE Internet Computing*, Vol. 15, no. 2, pp. 50-53, 2011.

[4] Lori M. Kaufman, "Data Security in the World of Cloud Computing, *IEEE Security & Privacy*, vol. 7, no. 4, pp. 61 -64, 2009.

[5] Yogita Gunjal, J. Rethna Virjil Jeny, "Data Security and Integrity of Cloud Storage in Cloud Computing, *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 2, Issue 4, April 2013.

[6] Wei-Tek Tsai, Xin Sun, Janaka Balasooriya, "Service-Oriented Cloud Computing Architecture, in *Proceedings of Seventh International Conference on Information Technology*, IEEE 2010.

[7] Gangolu Sreedevi, C. Rajendra, "ICCC: Information Correctness to the Customers in Cloud Data Storage, *International Journal of Advanced Research in Computer Engineering & Technology*, Volume 1, Issue 4, June 2012.

[8] Mohammad Sajid, Zahid Raza, Cloud Computing: Issues & Challenges, in *Proceedings of International Conference on Cloud, Big Data and Trust*, 2013, RGPV.

[9] Kalpana Batra, Ch. Sunitha, Sushil Kumar, "An Effective Data Storage Security Scheme for Cloud Computing, *International Journal of Innovative Research in Computer and Communication Engineering* Vol. 1, Issue 4, June 2013.

[10] Sherif El-etriby, Eman M. Mohamed, "Modern Encryption Techniques for Cloud Computing Randomness and Performance Testing, ICCIT 2012

[11] Qian Wang, Cong Wang, Kui Ren, Wenjing Lou, Jin Li, "Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing, *IEEE Transactions on Parallel and Distributed Systems*, Volume: 22 , Issue: 5 , May 2011.

[12] GaidaaSaeed Mahdi, "A Modification of TEA Block Cipher Algorithm for Data Security (MTEA), *Engg.& Tech. Journal*, vol 29, No.5, 2011.

Security Aspects of Virtualization in Cloud Computing

Archa Dharman, Henna Rose Babu and Jincy Varghese
Vidya Academy of Science & Technology
Thrissur- 680501, India

Sajay K R
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: sajay.k.r@vidyaacademy.ac.in

***Abstract*—In cloud computing, virtualization is the basis of delivering Infrastructure as a Service (IaaS) that separates data, network, applications and machines from hardware constraints. Although cloud computing has been a focused area of research in the last decade, research on cloud virtualization security has not been extensive. In this paper, different aspects of cloud virtualization security have been explored. Specifically, we have identified: i) security requirements for virtualization in cloud computing which can be used as a step towards securing virtual infrastructure of cloud, ii) attacks that can be launched on cloud virtual infrastructure, and iii) security solutions to secure the virtualization environment by overcoming the possible threats and attacks**

***Index Terms*—Virtualization, hardware, storage, software, virtual machine, attacks, hypervisor**

I. Introduction

In COMPUTING, a process of creating a illusion of something like computer hardware, operating system (OS), storage device, or computer network resources is virtualization. Cloud computing is one of the most useful technology that is been widely used all over the world. It generally provides on demand IT services and products. Virtualization plays a major role in cloud computing as it provides a virtual storage and computing services to the cloud clients which is only possible through virtualization. Cloud computing is a new business computing paradigm that is based on the concepts of virtualization, multi-tenancy, and shared infrastructure. This paper discusses about cloud computing, how virtualization is done in cloud computing, virtualization basic architecture, its advantages and effects. The main aim of virtualization is to manage the workload by transforming traditional computing to make it more scalable, efficient and economical. Virtualization can be applied to a wide range such as operating system virtualization, hardware-level virtualization and server virtualization. Virtualization technology is hardware reducing cost saving and energy saving technology that is rapidly transforming the fundamental way of computing.

II. Types of Virtualization

The different types of virtualisations are discussed in this section.

A. Hardware Virtualization

It is the most common type of virtualization and it provides advantages like optimum hardware utilization and application uptime. The basic idea is to combine many small physical servers into one large physical server, so that the processor can be used more effectively.

The hypervisor controls the processor, memory, and other components by allowing different OS to run on the same machine without the need for a source code. Hardware virtualization is further subdivided into the following types:

- Full virtualization:
  In this type of virtualisation, the complete simulation of the actual hardware takes place to allow software to run an unmodified guest OS.
- Para virtualization:
  In this type of virtualization, software unmodified runs in modified OS as a separate system.
- Partial virtualization:
  In this type of hardware virtualization, the software may need modification to run.

B. Storage Virtualization

In this type of virtualization, multiple network storage resources are present as a single storage device for easier and more efficient management of these resources. It provides various advantages as follows:

- Improved storage management in a heterogeneous IT environment
- Easy updates, better availability
- Reduced downtime
- Better storage utilization
- Automated management

C. Software Virtualization

It provides the ability to the main computer to run and create one or more virtual environments. It is used to enable a complete computer system in order to allow a guest OS to run. For instance letting Linux to run as a guest that is

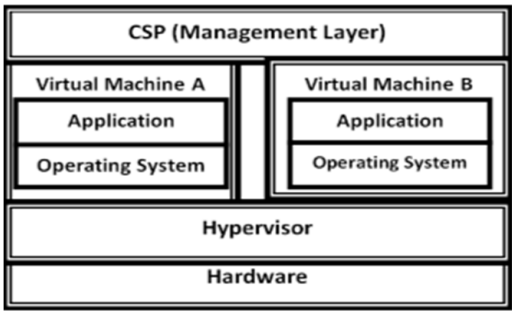

Fig. 1. Bare metal virtualization architecture

natively running a Microsoft Windows OS (or vice versa, running Windows as a guest on Linux).

The following are the different types of storage virtualisation:

- Operating system
- Application virtualization
- Service virtualization

III. Security Requirement of Virtualization

Different virtualization approaches can be applied to various system layers including hardware, desktop, operating system, software, memory, storage, data and network. Full virtualization is a form of hardware virtualization that involves complete abstraction of underlying hardware and provides better operational efficiency by putting more work load on each physical system. Full virtualization can be categorized into two forms:

- Bare metal virtualization
- Hosted virtualization.

Bare metal approach is mostly used for server virtualization in large computing systems like Cloud computing as it provides better performance, more robustness and agility. The architecture of bare metal based virtualization generally used in Cloud is shown in Fig. 1.

The unique characteristics of virtualization along with their benefits also have some drawbacks. Each component of virtualization needs to be secured from the possible threats. In general, before planning and implementing security of any system it is important to understand the security requirements of that environment. This section presents general requirements to prevent virtualization layers attacks in cloud.

A. Service Provider Requirements

A report by Alert Logic shows that 50 percent of Cloud users consider service provider security as a major threat. However, the impact of Cloud service provider on cloud virtualization security has also not been discussed comprehensively in literature. To secure the virtualization hardware, (Cloud) service provider must limit access of hardware resources to authorized person. Similarly, proper access control should be implemented in the management layer, so that each administrator has access only to its concerned data and software. The service provider also need to provide strong authentication mechanisms to users. Furthermore, security principles for the development of trusted computing system such as economy of mechanism, complete mediation, open design, principle of least privilege, psychological acceptability must also be followed by the service provider.

B. Hypervisor Requirements

Hypervisor provides the necessary resource management functions that enable sharing of hardware resources between the VMs. Hypervisor must maintain the isolation between VMs and support multiplexing of multiple VMs on single hardware platform. It must ensure that no application from any VM can directly take control of it as a host to modify the source code of hypervisor and other VMs in the network. Hypervisor should also monitor the guest OS and applications in VMs to detect any suspicious behavior Programs that control the hypervisor must be secured using similar practices used for security of programs running on servers. Similarly access to the hypervisor must be restricted. Other security measures to secure hypervisor include installing updates to the hypervisor, restricting administrator access to the hypervisors management interfaces and analyzing hypervisors logs to see if it is functioning properly.

C. Virtual Machine Requirements

Limit on VM resource usage has to be assigned so that malicious VMs can be restricted from consuming extra resources of the system. Moreover, isolation between virtual machines should be provided to ensure that they run independently from each other. To secure the guest OS running in virtual machines, best practices for the security of physical machines must be followed that include updating the OS regularly for patches and updates, using anti-virus software, securing internet and email and monitoring of guest OS regularly. Privileged VM (Dom0) is the first domain started by XEN hypervisor after boot. It is responsible for monitoring the communication between the remote users and guest VMs. Dom0 is also responsible for creating and destroying all guest VMs and providing device drivers to the guest VMs. Dom0 should boot the guest VMs without tampering them. The state of the VM saved as a disk file in Dom0 must remain confidential, and it must not be tampered.

D. Guest Image Requirements

Hypervisors use disk images (host files used as disk drive for guest OSs) to present guest OSs with virtual hard drives. Guest OS images can be moved and distributed easily, so they must be protected from unauthorized access, tampering and storage. To securely manage the guest OS images they must be examined and updated regularly according to the requirements. Unnecessary images must not be created and if any image is useless it must be removed from system. Whenever VM is migrated from one physical machine to another, images on previous disks should be completely removed. Similarly, data on old broken disks should also be removed before they

are discarded. Furthermore, backup of the virtual machines images must be maintained. VM checkpoint is a feature that allows the users to take snapshot of VM image in the persistent storage. Snapshot records the state of the running image that contains all components of the guest OS. Snapshot is generally captured as a difference between the image and the running state. The major function of checkpoint is to restore VM to its previous state if the VM enters any undesired state. However, the snapshot access should be given to authorized users and checkpoint must be used only to return VM to a stable and non-malicious state.

IV. Attacks on Virtualization

Each component of virtualization layer can act as an attack vector to launch multiple attacks on the system. Attacks that target different components of virtualization environment may result in security issues such as compromise of complete Cloud infrastructure, stealing of customer data and system hacking. This section discusses different attack scenarios at virtualization environment in Cloud.

A. Service Provider Attacks

If the attacker has physical access to the Cloud hardware, he may run malicious application or code in the system to damage the VMs by modifying their source code and changing their functionality. With the help of physical access to system, attackers can also launch cross VM side channel attacks. These attacks Security Aspects of Virtualization in Cloud Computing include CPU cache leakage to measure the load of other virtual web server on the network. Moreover, if access control is not implemented properly, different administrators such as network admin and virtualization admin might access the customer data that they are not authorized to access. These activities will result in security compromises such as loss of data confidentiality and unauthorized traffic monitoring. Service provider has to ensure that software deployed on Cloud are built using proper coding practices. Flawed coding can result in web application attacks such as SQL Injection, Cross Site Scripting, Denial of Service and Code Execution etc. Alert Logic report shows web application attacks to be the most common attacks on Cloud environment, impacting almost 52 percent customers.

B. Hypervisor Attacks

A Cloud customer can lease a guest VM to install a malicious guest OS, which attacks and compromises the hypervisor by changing its source code in order to gain access to the memory contents (data and code) of VMs present in the system. With more features in hypervisor its increased code size has resulted in design and implementation vulnerabilities. To control the complete virtualization environment malicious hypervisors such as BLUEPILL rootkit, Vitriol and SubVir and are installed on the fly, which give attacker the host privileges to modify and control VMs . This technique used by malicious software to take complete control of the underlying operating system by hiding itself from administrator and security software is called hyperjacking. Another attack in which program running in one VM can get root access to the host machine is called VM Escape. It is done by crashing the guest OS to get out of it and running an arbitrary code on the host OS. Therefore, such malicious VMs can take complete control of the host OS. Escaping the guest OS allows the VMs to interact with the hypervisor and provides them access to other guest OS on the system as well. Fig. 2 shows that the attacker from his virtual machine (VM 2) is able to escape his VM. VM 2 is used to compromise the hypervisor which is further used to launch attacks on other VMs (VM 1) in the system.

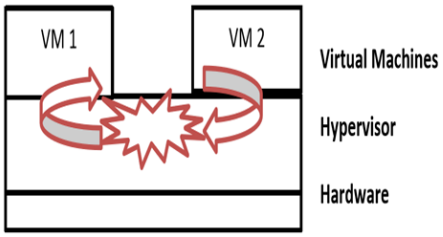

Fig. 2. VM Escape attack

C. Virtual Machine Attacks

Malicious programs in different virtual machines can achieve required access permissions to log keystrokes and screen updates across virtual terminals that can be exploited by attackers to gain sensitive information. If isolation is not properly implemented covert channels can be used for unauthorized communication with other VMs in the system. Attackers can use Trojans, malwares and botnets for traffic monitoring, stealing critical data, and tampering the functionality of guest OS. Conficker, Zeus botnet, command and control botnet communication activity are the examples of such attacks that result in data destruction, information gathering and creation of backdoors for attackers. Attacks through buggy software, viruses and worms can exploit the guest OS in VMs.

Furthermore, unpatched VM operating systems can be exploited by zero day attacks. The privileged host virtual machine Dom0 can be compromised by attacker to either tamper boot process of guest VMs or access all guest VMs including their memory, disk space and network traffic. By controlling Dom0 attacker can create too many virtual machines to consume all resources of the system or destroy any virtual machine containing important data by launching DOS attack at Cloud. Furthermore, the saved state of guest virtual machine as a disk file appears in plaintext to Dom0. Attacker can compromise the integrity and confidentiality of the saved VM state and when restored VM may not function as desired.

D. Guest Image Attacks

Unnecessary guest OS images in Cloud can result in different security issues if the security of each image is not maintained. If a malicious guest OS image is migrated to

another host, it can compromise the other system as well. Furthermore, creating too many images and keeping unnecessary images can consume resources of the system which can be used as a potential attack vector by attacker to compromise the system [2]. When VMs are moved from one physical machine to other, data of VM images might still exist on previous storage disks that attacker can access. Similarly, attackers might also recover some data from old broken disks. The security of image backup is also an issue. By gaining access to the backup images attacker can extract all information and data. Attacker can access VM checkpoint present in the disk that contain VM physical memory contents and can expose sensitive information of VM state. A new checkpoint can be created by attacker and loaded in system to take VM to any state desired by attacker. If all the checkpoints in storage are accessed, information about previous VM states can be obtained.

V. Security Solutions for Virtualization

To cater the attacks on virtualization environment different security solutions have been proposed in literature. This section discusses those security solutions for each component of virtualization architecture. By implementing these security solutions the attacks discussed in section 3 can be mitigated or at least the impact of those attacks on virtualization environment can be minimized.

A. Service Provider Security

Unauthorized person should not have physical access to the virtualization hardware of the system. In order to protect VMs from unauthorized access by Cloud administrators, each VM can be assigned access control that can only be set through Hypervisor. The three core principles of access control namely identification, authentication and authorization will restrict admin access from unauthorized data and system components. Moreover, if any administrator is involved in security compromise, access control implemented in Cloud can help identify that person. Web application attacks can be prevented by installing an application layer firewall infront of web facing applications and by having the customer code reviewed for common vulnerabilities. An online identity management community OpenID has been integrated with an open source Cloud platform OpenStack to provide identity management in Cloud. Sandra R. et al. proposed an architecture using SELinux, XEN, IPsec as tools to enforce Mandatory Access Control (MAC) policies at VM, OS and network layers. These MAC policies control the communication between VMs based on application templates that can be configured by administrators dynamically. Furthermore, the security requirements of virtualized environment differ from that of physical system, Cloud service provider must make sure that the security tools for vulnerability assessment also include the virtualization tools used.

B. Hypervisor Security

Hypersafe is a system that maintains code integrity of the Hypervisor. It extends the hypervisor implementation and prevents its code modification by locking down the write-protected memory pages. It secures the Hypervisor against the control-flow hijacking attacks by protecting its code from unauthorized access [15]. VM Escape attack can only be executed through a local physical environment. Therefore, the physical Cloud environment must be prevented from insider attacks. The interaction between guest machines and host OS must also be properly configured. In order to stop one VM from affecting or communicating with other VMs isolation must be properly implemented and maintained by hypervisor. Moreover, further possible attack vectors on hypervisors can be reduced by hardening the hypervisor. These techniques include separating the duties of administrative functions, restricting the hypervisors administrator access to modify, create or delete hypervisor logs, and monitoring the hypervisor logs regularly

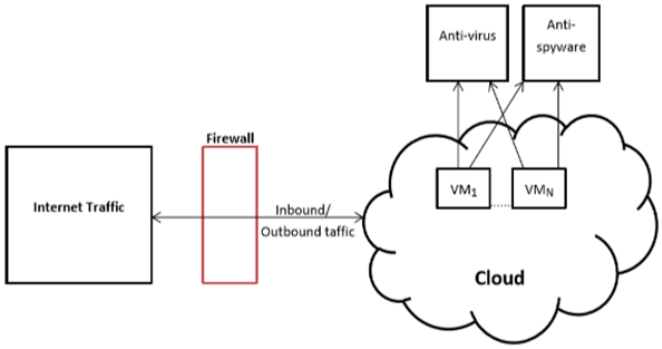

Fig. 3. VM security

C. Virtual Machine Security

Administrator must deploy a software or application that stops VMs from using extra resources unless authorized. Moreover, a light weight process must run on a virtual machine that collects logs from the VMs and monitors them in real time to fix any tampering of VMs. The guest OS and applications running on it must be hardened by using best security practices. These practices include installing security software such as anti-viruses, anti-spyware, firewall, Host Intrusion Prevention System (HIPS), web application protection, and log monitoring in guest OS [4]. Protection of VMs by different security practices is shown in Fig. 3. To identify the faults in guest OS Dan P. et al. proposed a system called Vigilant. It utilizes virtualization and machine learning methods to monitor VMs through hypervisor without putting any monitoring agent in VMs (outof-band detection). Flavio L. et al. proposed Advanced Cloud Protection System (ACPS) that monitors and protects the integrity of OS in guest VMs. The periodic monitoring of executable system files is done to check the behavior of Cloud components. It uses virtual introspection techniques to deploy guest monitoring machine in system without being noticed by attacker on guest VM. Hence any suspicious activity on the guest OS can be blocked. To protect the newly created virtual machines for users (guest

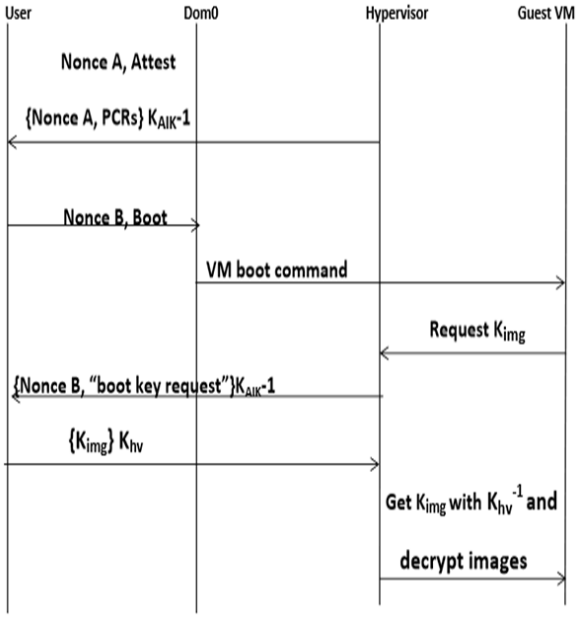

Fig. 4. Secure VM boot protocol

VMs) from compromised privileged virtual machine Dom0, a protocol is designed by Jinzhu Kong. Hypervisor generates a pair of secret keys, Kernel and the initrd image are kept encrypted all the time with the secret key Kimg. First the user attests the Cloud server through Trusted Platform Module (TPM), if attestation succeeds then user sends a boot request to the Dom0 which then boots the guest domain. The guest VM executes the wrapping code and requests Hypervisor to decrypt kernel and initrd images. Hypervisor encrypts this request with its private key and asks user for key to decrypt kernel so that a VM can be created. The user sends private key Kimg encrypted under the public key of Hypervisor. Hypervisor decrypts the user message, and the private key Kimg is used to decrypt the kernel, initrd images and to launch the guest virtual machine. In this way the newly created VM is secured from compromised Dom0. The complete workflow is shown in Fig. 4. To avoid the VM storage attacks, before saving the state of the virtual machine in Dom0 its encryption can be done using AES-256, where key can be any random initialization vector. The hash of the encrypted state can be taken using MD5. When the virtual machines are to be restored, the new hash can be taken to verify the integrity of saved virtual machine. If the hash of the restored state and hash of the saved state match it means that the virtual machine state is not altered. Fig. 5 shows the secure storage of saved VM state.

D. Guest Image Security

Organizations using virtualization must have a policy to manage the creation, usage, storage and deletion of images. Image files must be scanned for the detecting viruses, worms, spyware and rootkits that hide themselves from security software running in guest OS. J. Wei et al. proposed an

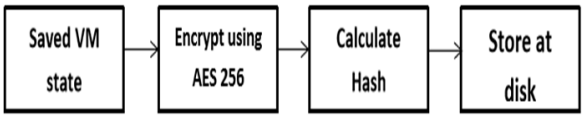

Fig. 5. Securing the saved VM state

image management system to efficiently manage images in Cloud and detect security violations in images. It proposes the use of filters, virus scanners and rootkit detectors to provide protection against potentially compromised images. Nuwa is a tool designed to apply efficient patching to VM images in Cloud. By analyzing patches, Nuwa rewrites the patching scripts so that they can be applied offline. As a result, the installation scripts for online patching can be applied to images when they are offline. When VMs are to be migrated from one physical machine to another, Cloud admin must recheck and ensure that all data is removed from previous or broken disks. To protect the backup VM images cryptographic techniques such as encryption may be employed to encrypt all backup images. If any VM is deleted then its backup must also be removed from system. Furthermore, to protect VM images from storage attacks, Cloud provider must encrypt the complete VM images when not in use [3]. Checkpoint attacks can be prevented by encrypting the checkpoint files. Another mechanism to provide security to Checkpoints is SPARC. SPARC is a mechanism designed to deal with security and privacy issues resulting from VM checkpoint. SPARC enables users to select applications that they want to checkpoint so sensitive applications and processes cant be checkpointed. Table 1 shows the summary of different security aspects of virtualization discussed in the paper.

VI. Conclusion

The security of cloud cannot be maintained unless its virtualization environment is secured. Although different virtualization approaches exist, bare metal virtualization approach is commonly used in large computing systems such as Cloud for server virtualization. This paper presents general architecture of bare metal virtualization and covers security aspects of its different components. Cloud virtualization environment can be compromised by different attacks at service provider, hypervisor, virtual machines, guest operating system and disk images. The attack scenarios at these components are discussed in the paper. To provide security to the virtualization environment, general requirements for virtualization security and different existing security schemes that provide security to virtualization environment have also been discussed. Therefore, the holistic picture of virtualization security in Cloud is provided through structured analysis in which security requirements, attacks and solutions correspond to each other. Addressing these security aspects will lead towards more extensive research on secure Cloud virtualization environment. In future, an assessment criteria needs to be proposed by which we can analyze the effectiveness of security solutions of virtualization against the

specific attacks.

REFERENCES

[1] Dave Shackleford, *Virtualization Security: Protecting Virtualized Environments*, Sybex, 1st edition, November 2012.

[2] Matthew Portnoy, *Virtualization Essentials*, Sybex, 2nd edition, August 2016. by Matthew Portnoy

[3] Matthew Portnoy, *Virtualization Essentials*, 1st Edition, 2012 by Sybex, John Wiley & Sons, 2012.

[4] Muhammad KazimRahat MasoodMuhammad Awais ShibliAbdul Ghafoor Abbasi, "Security Aspects of Virtualization in Cloud Computing", in *Computer Information Systems and Industrial Management*, CISIM 2013.

[5] Deb Shinder (2008), "Security Through Virtualization" [Online]. Available: http://techgenix.com/security-through-virtualization/.

[6] Kaushik Pal(2015), "10 Ways Virtualization Can Improve Security" [Online]. Available: https://www.techopedia.com /2/31007/trends/virtualization/10-ways-virtualization-can-improve-security

[7] Bill Kleyman (2015), "Five Security Best Practices for Cloud and Virtualization Platforms" [Online]. Available: https//www.datacenterknowledge.com/archives/2015/11/09/ five-security-best-practices-for-cloud-and-virtualization-platforms.

Cyber Crimes and Cyber Laws

Arya A, Leo Joy and Neeha Maria M
Vidya Academy of Science & Technology
Thrissur - 680501, India

Salkala K S
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: salkala@vidyaacademy.ac.in)

***Abstract*—As we all know that this is the era where most of the things are done usually over the internet starting from online dealing to the online transaction. Since the web is considered as worldwide stage, anyone can access the resources of the internet from anywhere. The internet technology has been using by the few people for criminal activities like unauthorized access to others network, scams etc. These criminal activities or the offense/crime related to the internet is termed as cyber crime. In order to stop or to punish the cyber criminals the term Cyber Law was introduced. We can define cyber law as it is the part of the legal systems that deals with the Internet, cyberspace, and with the legal issues. It covers a broad area, encompassing many subtopics as well as freedom of expressions, access to and utilization of the Internet, and online security or online privacy. Generically, it is alluded as the law of the web.**

I. Cyber Crime

Sussman and Heuston first proposed the term Cyber Crime in the year 1995. In simple term we can describe cyber crime are the offences or crimes that takes place over electronic communications or information systems. Due of the development of the internet, the volumes of the cyber-crime activities are also increasing because when committing a crime there is no longer a need for the physical present of the criminal.

A. Evolution of Cyber Crime

The following table summarises the timeline of the evolution of cyber crimes.

TABLE I
Evolution of Cyber Crime

1997	Cyber crimes and viruses initiated, that includes Morris code worm and other.
2004	Malicious code, Torjan, Advanced worm, etc.
2007	Identifying thief, Phishing, etc.
2010	DNS attack, rise of Botnets, SQL attacks, etc.
2013	Social Engineering, DOS attack, BotNets, Malicious Emails, Ransomware attack, etc
Present	Banking Malware, Keylogger, Bitcoin, Wallet, Phone hijacking, Android hack, Cyber warfare, etc.

B. Characteristics/Constraints of Cyber Crime

Stealing information from computers has proven to be accurate, easy, reliable, cheap and less risky in terms of detection. Some more constraints/characteristics of cyber crime are:

- Low marginal cost of online activity due to global reach.
- Lower risk of getting caught.
- Catching by law and enforcement agency is less effective and more expensive.
- New opportunity to do legal acts using technical architecture.
- Official investigation and criminal prosecution is rare; not very effective sentences.
- No concrete regulatory measure.
- Lack of reporting and standards
- Difficulty in identification
- Limited media coverage.
- Corporate cyber crimes are done collectively and not by individual persons.

C. Classification of Cyber crime

Cyber crime can be classified into four major categories.

1) Cyber crime against individual:

1) Email Spoofing:
   This technique is a forgery of an email header. This means the message appears to have received from someone or somewhere other than the genuine or actual source. People are probably going to open an electronic mail or an email when they think that the email has been sent by a legitimate source.
2) Spamming:
   It is also known as junk email. It is unsought mass message sent through email. The recipients email addresses are obtained by spam bots, which are automated programs that crawls the internet in search of email addresses. The spammers use spam bots to create email distribution lists. With the expectations of receiving a few number of respond a spammer typically sends an email to millions of email addresses.

3) Cyber defamation:
   It means, harm that is brought on the reputation of an individual in the eyes of other individual through the cyber space. The purpose of making defamatory statement is to bring down the reputation of the individual.
4) IRC Crime (Internet Relay Chat):
   IRC servers allow the people around the world to come together under a single platform which is something called as rooms and that chat to each other.
5) Phishing:
   The attacker tries to gain information such as login information or accounts information by masquerading as a reputable individual or entity in various communication channel or in email.

2) Cyber Crime against Property: This type of crimes includes vandalism of computers, Intellectuals (copy right, patented, trademark, etc) property crimes, online threatening, etc. Intellectual property crime includes:

1) Software privacy:
   Copying of software unauthorizedly.
2) Copyright infringement:
   Infringement of an individual or organizations copyright. Simply, the using of copyright materials unauthorizedly such as music, software, text, etc
3) Trademark Infringement:
   Using of a service mark or trademark unauthorizedly.

3) Cyber Crime against Organization:

1) Unauthorized changing or deleting of data:
   Reading or copying of confidential information unauthorizedly, but the data are neither being change nor deleted.
2) DOS attack (Denial Of Service):
   The attacker floods the servers, systems or networks with traffic in order to overwhelm the victim resources and make it infeasible or difficult for the users to use them.
3) Email bombing:
   It is a type of Net Abuse, where huge numbers of emails are sent to an email address in order to overflow or flood the mailbox with mails or to flood the server where the email address is.
4) Salami attack:
   It also known as Salami slicing. The attackers use an online database in order to seize the customers information like bank details, credit card details, etc. Attacker deduces very little amount from every account over a period of time.

4) Cyber Crime against Society:

1) Forgery:
   Making of false document, signature, currency, revenue stamp, etc.
2) Web jacking:
   It is derived from hi jacking. The attacker creates a fake website and when the victim opens the link a new page appears with the message and they need to click another link. If the victim clicks the link that looks real he will redirected to a fake page. These type of attacks are done to get entrance or to get access and controls the site of another.

D. Common Tools and Techniques of Cyber Crime

Unauthorized access is the main tool used by criminals. Unauthorized access means any kind of access without the permission of rightful owner or in charge of the computer, computer system or computer network. Following are the common techniques used for unauthorized access:

1) Packet Sniffing:
   Packet sniffer is a utility that has been used since the original release of Ethernet. Packet sniffing allows individuals to capture data as it is transmitted over a network. This technique is used by network professionals to diagnose network issues, and by malicious users to capture unencrypted data, like passwords and usernames. If this information is captured in transit, a user can gain access to a system or network.
2) Password Cracking:
   A password is a type of authentication. It is a secret word or phrase that a user must know in order to gain access. All systems cache passwords in memory during login session. Therefore, if a hacker can gain access to all memory on the system, he can likely sift the memory for passwords. Likewise, hackers can frequently sift page files for passwords. To crack a password means to decrypt a password, or to bypass a protection scheme. Another form of password cracking attack is brute force attack. In this form of attack, all possible combinations of letters, numbers and symbols are tried out one by one, till the password is found out. Brute force attacks take much longer than other attacks.
3) Buffer Overflow:
   A buffer is a temporary area for data storage. When more data (than was originally allocated to be stored) gets placed by a program or system process, the extra data overflows. It causes some of that data to leak out into other buffers, which can corrupt or overwrite whatever data they were holding. In a buffer overflow attack, the extra data sometimes holds specific instructions for actions intended by a hacker or malicious user; for example, the data could trigger a response that damages files, changes data or unveils private information.

E. Computer Crime

Types of malicious software:

1) Virus:
   Infect computers by watching users action and accordingly strikes.
2) Worms:
   Worms are used as a conduit by attackers to control the victim computer and install their own copies to get information.
3) Trojan:
   This is installed when users is downloading or installing any program. The malware performs unexpected actions

without the knowledge of user like external access to the computer.

4) Spyware:
Acts like a spy and transmits the information such bank details, credit card information etc. Key logging software gathers all the information entered by using the keyboard and then sends back to the criminals.

TABLE II
PREVENTIVE MEASURES TO OPPOSE CYBER CRIME

Adopt computer security	Avail new sophisticated products and advice for computer crime prevention which is available free or paid in the market.
Educate Children	Children should be taught about the child pornography crime used by criminals and how to avoid that.
Install Original Software	As they contain many security measures. Pirated softwares do not contain many security abilities which exist in the original software.
Attachments	Avoid opening attachments or e-mails which were not expecting and have come from unknown source or person.

F. Cyber Criminals

Some known cyber criminals are:

1) Kids:
Kids take pride in hacking into a computer system or a web site. They also commit cyber crimes innocently without knowing implications.
2) Organized hacktivists:
Social activism and religious activism attacks prominent web sites for political reasons.
3) Disgruntled employees:
Instead of going on strike previously they commit computer related crimes due to automation process and this brings entire system to collapse.
4) Professional hackers:
Business Organizations store all information in their computers and the employees of rival organizations hack or steal the secrets for their benefit.

G. Reason for Growth of Cyber Crime

1) Motivation:
Intellectual challenge of mastering complex system was the motivation in the past for criminals, but presently criminals are driven by greed, lust, power, revenge, adventure.
2) Opportunities:
Growth of computing abilities in banking, stock exchange, air traffic control, telephones, electric power, health welfare institution and education, has though brought down the cost leading to revolutionary changes in commerce, communications, entertainment and education, and is providing more criminal opportunities owing to few vulnerabilities that exist in information technology
3) Ignorance:
Due to constant decrease in the prices of computers, more and more people are using computers but are unaware of potential threats from computer crime since many people do not possess sufficient technical skills to safeguard themselves from computer crimes.
4) Broadband:
Further, high speed internet connections which have now become easily available can handle high volume of network traffic and act as catalyst for computer crime.
5) Poor/limited response from Law enforcing agencies:
Many developing countries lack appropriate law to tackle the cyber crime attackers.

II. CYBER LAW

Cyber law plays a very important role in this new epoch of technology. It is important as it is concerned to almost all aspects of aspects of activities and transactions that take place either on the internet or other communication devices.

A. Cyber Law Awareness Program

Once should have the following knowledge in order to stay aware about the cyber crime;

1) One should read the cyber law thoroughly
2) Basic knowledge of internet and internets securityRead cyber crimes cases. By reading those cases one can be aware from such crimes. Trusted application from trusted site can be used for protection of ones sensitive information or data.
3) Technologys impact on crime

B. Cyber Laws in India

Following are the sections under IT Act 2000:

1) Section 65:
Temping with the computers source document : Whenever intentionally or knowingly destroy, conceal or change any computers source code that is used for a computer, computer program and computer system or computer network.
Punishment: Any person who involves in such crimes could be sentenced upto 3 years imprisonment or with a fine of Rs. 2 lakhs or with both.
2) Section 66:
Hacking with computer system, data alteration, etc. : Whenever with the purpose or intention to cause any loss, damage or to destroy, delete or to alter any information that resides in a public or any persons computer. Diminish its utility values or affects it injuriously by any means commits hacking.

Punishment: Any person who involves in such crimes could be sentenced upto 3 years imprisonment, or with a fine that may extend upto 2 lakhs rupees, or both.

C. A Few Important Sections One Should Know

TABLE III
IMPORTANT SECTIONS OF IT ACT

OFFENCES	SECTION UNDER IT ACT 2000
Damage to computer, Computer system, etc	Section 43
Power to issue direction for blocking from public access of any information through any computer's resources.	Section 69A
ower to authorize to collect traffic information or data and to monitor through any computer's resources for cyber security.	Section 69B
Unauthorized access to protected system.	Section 70
Sending threatening message by email.	Section 503 IPC
Email spoofing	Section 463 IPC
Web jacking	Section 463 IPC
Online sale of arms	Arm Act

D. Law Enforcement Agencies

Cyber crime is often difficult for Law Enforcement Agencies to investigate and control since there exists many constraints before them. The constraints and suggestive measures to counter them are mentioned in Table.

TABLE IV
CONSTRAINTS AND SUGGESTED MEASURES FOR LAW ENFORCEMENT AGENCIES

Constraints	*Suggested Counter Measures*
Lack of Funds (whereas law breakers have enough funds for best hardware and softwares)	• National Repository to be established for investigation into cyber crime. • More funds for training computer forensic personnel. • Sufficient funds to be kept aside every year for upgrading security system in future.
Lack of Latest Technology and Good Equipments	• Developing investigations tools for cyber crime in advance e-mail tracking. • Investigators to be provided with latest equipments • Easy access to technology required, to conduct computer investigation • Use of specialized software and training.
Lack of Training	Investigators to be continuously trained for (a) proficiency in investigating cyber crimes (b) projecting the future complexes of cyber crimes (c) locating computer based evidences (d) understanding cyber crime legal challenges [8]
Documentation and Procedures are not adequately defined	• Describing advance search and seizure procedure in documentation to handle high volume crimes. • Since cyber crimes are diverse in nature different kinds of documentations are required for security system. • Reporting standards to be developed for investigating the crime.
Lacking Forensic Support	• Computerized forensic support required for making strong forensic imaging and verification. In order to follow 'footprints' both on the computer and on the Internet • Creating forensic software and using high storage hard drives and good equipments to maintain high standards for recovery and preservation of evidence [9].

III. PROTECTION FROM CYBER CRIME

Dr. Les Labuschagne from the California Berkley University suggests two approaches: proactive and reactive. Most organizations adopt a reactive approach to information security. The vulnerability of systems is usually evaluated after an attack takes place, resulting in money spent on fixing the security holes and recovering from the data and business loss. This is the least effective, and more expensive approach. The proactive approach said to demonstrate organizations that try to locate security holes before the hackers do. The proactive approach is sometimes called ethical hacking. There are several websites, providing information for children as well as for parents:

1) Teach kids to be a good cyber citizen using Cybercitizenship rules at http:// www. cybercitizenship.org
2) Let kids take the CyberSpacers' oath and join the Super Cyber Team, from CyberSpacers website at www.cyberspacers.com. The site includes online quizzes, comics, games and contests
3) Show kids how to use the Internet safely and responsibly, and let them find out what happened to a young hacker, available from Cyberethics for Kids website at www.cybercrime.gov /rules/
4) The FBI site includes links to Internet Law Enforcement Stories and A Parent's Guide to Internet Safety Tips for Kids at http://www.fbi.gov/fbikids.html

A. Safety in Cyberspace

If possible always use a strong password and enable two-step or twostep authentication in the webmail. It is important in order to make your webmail or your social media account secured.

Two-step authentication is an additional layer of security that requires your username and the password also a verification code that is sent via SMS to the registered phone number. A hacker may crack your password but without the temporary and unique verification code should not be able to access your account.

Even with the knowledge of attacks prevention, technical abilities to protect the systems and laws available, computer crime is still spreading. Why do we see such a low number of attacks reports? Companies just don't want the publicity.

When something is going wrong with a company, the last thing they want is for everybody to know about it. A successful attack may challenge other hackers to repeat the crime. Bad publicity can seriously undermine the image and reputation of the company, as well as public trust.

IV. Conclusion

The rise and proliferation of newly developed technologies begin star to operate many cyber crimes in recent years. Cyber crime has become great threats to mankind. Protection against cyber crime is a vital part for social, cultural and security aspect of a country. The Government of India has enacted IT Act, 2000 to deal with cyber crimes. The Act further revised the IPC, 1860, the IEA (Indian Evidence Act), 1872, the Banker's Books Evidence Act 1891 and the Reserve Bank of India Act, 1934. Any part of the world cyber crime could be originated passing national boundaries over the internet creating both technical and legal complexities of investigating and prosecuting these crimes. The international harmonizing efforts, coordination and co-operation among various nations are required to take action towards the cyber crimes.

Our main purpose of writing this paper is to spread the content of cyber crime among the common people. At the end of this paper we want to say cyber crimes can never be acknowledged. If anyone falls in the prey of cyber attack, please come forward and register a case in your nearest police station. If the criminals wont get punishment for their deed, they will never stop.

References

[1] Foote D. (2002, March). "Good Ethics at Work Lie in the Hiring", *Computerworld*. [Online] Available: https://www.computerworld.com/article/2587431/good-ethics--at-work-lie--in-the-hiring.html.

[2] Harvey B (2004). "Computer hacking and ethics", University of California, Berkeley. [Online] Available: http://www.cs.berkeley.edu/ bh/hackers.html.

[3] *Internet Stuff*(2004, May, 25). "2004 E-Crime Watch Survey" [Online] Available:http://www.cert.org/about/ecrime.html.

[4] *Internet stuff* (2004). "What is cyber crime" [Online] Available: http://www.cybercitizenship.org/crime/crime.html.

[5] Khalid A. (2004, March 5). "Cyber crime: Business and the law on different pages" [Online] Available: http://www.niser.org.my/news/200403-05-01.html

[6] Labuschagne L. (2000, July). "Evaluation criteria", Rand Afrikaans University. [Online] Available: http://csweb.rau.ac.za/staff/labuschagne/research/articles/eth-hac.pdf.

[7] McCullagh D. (2003, August, 1). "Hackers get lesson in the law", Cnet-News. [Online] Available: http://news.com.com/2100-1009 5058918.html

Proxy Server: Wireless Sensor Network and SQL Injection Attack

Arya S A, Reshma V S and Susmitha P N
Vidya Academy of Science & Technology
Thrissur - 680501

Manesh D
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: manesh.d@vidyaacademy.ac.in)

***Abstract*—A proxy server is a web server that caches internet resources for re-use by a set of client machines. Proxy server provides various web services and it improves the quality of web service and also provides security. In this paper we investigate proxy server chained http network for increase web service quality and the interconnection between wireless sensor networks and the internet enabled by proxy server and cloud gateway. Then the SQL injection attack and prevention technologies . SQL injection defence model is established according to the detection processes, which is effective against SQL injection vulnerabilities.**

***Index Terms*—Proxy server, wireless sensor network, SQL injection attack**

I. Introduction

In a computer network, a proxy server is any computer system offering a service that acts as an intermediary between the two communicating parties, the client and the server. Proxy server provides various functionalities. Especially, to improve the quality of web service by using web cache and load balancing and it locate various placements. For this reason, users are provided service along many proxy services. This is leads to falling web performance, because of increasing HTTP message processing time and network latency. In web services the NATO (The North Atlantic Treaty Organization, also called the North Atlantic Alliance, is an intergovernmental military alliance between 29 North American and European countries.) has identified Web services as the key enabling technology for achieving application interoperability when interconnecting heterogeneous systems.

Sensor networks offer a unique service for applications in many fields like industrial automation, scientific data collection, logistics, home automation, or automotive industry. Sensor networks provide different characteristics like safe data transmission, high data rate or long batteries like span.

Wireless sensor network consists of a large number of micro sensor nodes and it is capable of real-time monitoring, sensing and collecting all kinds of information in the network distribution area. WSN used in smart home, environmental monitoring and other fields. Wireless sensor network cannot complete an application independently. The interconnection of WSN and internet control and manage the wireless sensor network devices in any place.

SQL Injection attack (SQLIA) is one of the techniques used to attack databases through a website. This attack tries to gain access to sensitive data directly by injecting malicious SQL codes through web application.SQL injection was an attack in which malicious code was embedded in strings that were later passed to database backend for parsing and execution. The malicious data produced database query results and acquired sensitive information, such as account credentials or internal business data. This paper studies a reverse proxy with an intrusion and prevention mechanism built in against web attacks like SQLIA. This method offers protection over servers residing in internal network while providing services to external client.SQL injection vulnerability allowed an attacker to a web application's underlying database and destroyed functionality or confidentiality. The detecting methods not only validate input values but also use type-safe SQL parameters, which is effective against SQL injection vulnerabilities.

II. Proxy Server

A proxy server is also known as a proxy or application-level gateway. A proxy server is any computer system offering a service that acts as an intermediary between the two communicating parties, the client and the server. How a proxy server is work for example, When a proxy server receives a request for an Internet resource (such as a Web page), it looks in its local cache of previously pages. If it finds the page, it returns it to the user without needing to forward the request to the Internet. If the page is not in the cache, the proxy server, acting as a client on behalf of the user, uses one of its own IP addresses to request the page from the server out on the Internet. When the page is returned, the proxy server relates it to the original request and forwards it on to the user.

A. Proxy Server Bypassing with TCP Connection

The proxy server bypass TCP connection by writing own IP and Port address via header of HTTP request message. Figure 2 represents an example of the chained HTTP proxy network

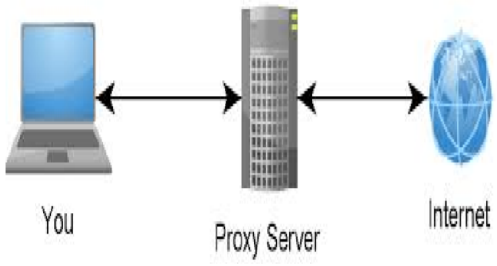

Fig. 1. proxy server

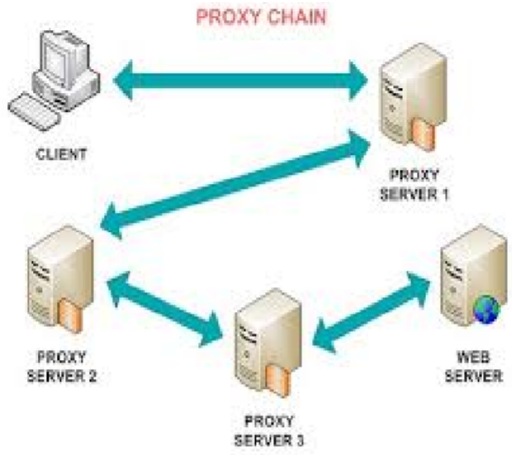

Fig. 2. Chained proxy network architecture

architecture. Proxy server by-pass connection configuration procedure starts that the client sends HTTP request message along proxy chain constituting proxy server A, B, C, D. Receiving HTTP request message from the client, Proxy A creates via header in the message and write own IP and Port address at via header. After creating and writing at the Via header, proxy A forward the modified HTTP message to proxy B. Receiving the messages, Proxy B additionally write IP and Port address at Via header. This procedure is repeated and then final modified HTTP request message came to proxy D. Proxy D establishes proxy server by-pass TCP connection with proxy A using Via header first written IP and Port address among all IP and Port address at Via header.

Then proxy D request web page to Server by removing via header, which makes origin HTTP request message. Server that received HTTP request message makes HTTP response message for web page. And the server sends a message to proxy D. It is not that the proxy D forwards HTTP response message to neighbour proxy C included proxy chain, but that the proxy D forwards HTTP response message to proxy A through by-pass TCP connection. Client receives the web page from the server by forwarding HTTP response message.

HTTP is the set of rules for transferring files (text, graphic image, sound, video, and other multimedia file) on the World Wide Web. It also defines how messages are formatted and transmitted and what actions Web servers and browsers should take. Web services are open standard (XML, SOAP, HTTP, etc.) based web applications that interact with other web applications for the purpose of exchanging data. Web services can convert your existing applications into web applications.

The performance evaluation and analysis of proxy server are classified into two types Simulation Parameter and Simulation Result and Analysis. In Simulation Parameter, we perform network simulation for the performance evaluation and analysis about the proposed scheme. Simulation scenario is that client requests web page to server along 5 proxy servers constituting proxy chain. And the client receives a web page from the server. In Simulation Result and Analysis, the proposed scheme improves page loading time by reducing HTTP message processing time and network latency. Now the most popular proxy server used today is a Web proxy, and it is used to filter contents and allow anonymous browsing.

III. Web Services

Web services are XML-based information exchange systems that use the Internet for direct application-to-application interaction. This technology has communication overhead, leading to a need to investigate ways to optimize its resource use when applying the technology in limited capacity networks. To reduce the overhead of Web services we can attempt to reduce the overhead of SOAP (originally Simple Object Access Protocol is a messaging protocol specification for exchanging structured information in the implementation of web services in computer networks), the Web services messaging standard.

In this paper we investigate the use of alternative transport protocols for Web services in military networks. The alternatives transport protocols are: TCP, UDP, SCTP, and AMQP.

A. Test Framework for SOAP

SPADE is a test framework that we have developed for performing SOAP analysis in dynamic environments. The framework consists of the following components:

1) A web service client plug-in:
   Requests a message from the service of a given payload size.
2) A network emulator:
   We use an IBX-200 embedded computer running Ubuntu 10.04 LTS and netem for emulating network properties such as bandwidth, delay, loss, packet reordering and duplication.
3) A request generator:
   Configures the network emulator, generates network traffic using the Web service and client, and measures the performance.
4) A proxy:
   The proxy translates between COTS Web services communication over HTTP/TCP and other protocols. It also supports data compression.
5) A Graphical User Interface (GUI):
   For generating requests and controlling the above components.

B. The Transport Protocols

We list below the various transport protocols.

1) TCP
   The Transmission Control Protocol (TCP) is one of the core transport protocols of the Internet Protocol Suite. It is connection-oriented and provides end-toend reliability. Furthermore, it is the most widespread transport protocol in use for Web services, as SOAP over HTTP/TCP is the preferred binding in most deployments.
2) UDP
   The User Datagram Protocol (UDP) is another core transport protocol of the Internet Protocol Suite, but unlike TCP it uses a simple transmission model without mechanisms providing flow control, packet ordering, or integrity of the messages.
3) SCTP
   The Stream Control Transmission Protocol (SCTP) was originally designed as a protocol for telephony signalling over IP networks. It offers functionality from both TCP and UDP, in that it is messageoriented like UDP but ensures reliable, in-sequence transport of messages with congestion control like TCP. SCTP has been implemented for all major operating systems. The two most important enhancements over the traditional transport protocols are the end-to-host capabilities (called multi-homing) and multi-streaming.
4) AMQP
   The Advanced Message Queuing Protocol (AMQP) is an emerging standard for business messaging that is currently in widespread use in the financial sector. AMQP is a complete messaging middleware that can utilize different transport protocols. Current implementations support both the request/response and the publish/subscribe communication paradigms. The main benefit of this middleware is its reliability when facing network disruptions, since it employs a broker-based architecture with store-andforward capabilities. This means that AMQP could be suitable for use in disruptive environments, provided the overhead is not too large. SPADE implements support for AMQP over TCP enabling evaluation of the protocol using the request/response paradigm, which is suitable for conveying traffic between our request/response Web service client and service. We used the RabbitMQ AMQP implementation in our framework.

We have used SPADE to evaluate the four transport protocols mentioned above.

IV. Interconnection between Wireless Sensor Network and Internet with Proxy Server and Cloud Gateway

There are several solutions to realize the interconnection between the wireless sensor network and internet. One solution is to implement the TCP/IP protocol stack on top of the stack of the sensor network. The translation of the packets from the sensor network to the TCP/IP-based internet and vice versa is realized in the application layer on a special gateway. This gateway performs a logical mapping from the sensor address to an IP address for each node. The sensor network stack keeps unchanged but each node is still accessible by an IP address. This IP address is said as Virtual IP address. Here the virtual IP gateway is used, for keeping the protocol stack of both the networks. It was implemented with help of a transparent proxy server. The transparent proxy is used to redirect IP packets which are accessed to virtual IP addresses, to the responsible gateway application. The XMLbased protocol is used to realize the communication between the gateway applications and the web server. A web server is used for generating HTML pages for the user. Using the web server, user can check the status of the whole sensor network and can perform different operations. ZigBee is a protocol stack for wireless sensor networks based on the IEEE 802.15.4 standard which defines the physical layer and the medium access control layer of the protocol stack and above these two layers the ZigBee stack is implemented. ZigBee network works independently of the gateway.

Wireless sensor network can connect with internet through cloud gateway. It provides the efficient service for the user with the use of cloud platform computing ability, storage ability and the ability of information service.

A. Fundamental Techniques

WSN can be connected in different ways. They are application-level gateway, Delay-tolerant network, Overlay network, and Virtual IP Bridge.

1) Application-level gateway:
   Gateway is used to connect WSN with internet and it is free to choose the communication protocol. Once the proxy server fails, all network communication will fail.
2) Delay-tolerant network (DTN):
   It simplifies the integration between heterogeneous wireless sensor network and increases the communication overhead.
3) Overlay network:
   There are two types of overlay network, they are Internet overlay WSN and WSN overlay internet. Here protocol stack on the WSN allows users to directly access the sensor node has an IP address and increases the system overhead.
4) Virtual-IP Bridge (VIP):
   The mapping between the node and IP addresses stored in the VIP bridge. It realizes the direct communication of the user and sensor node through the virtual IP address.

B. System Model Architecture

System model architecture has wireless sensor network (WSN), cloud gateway, and cloud server.

1) Wireless Sensor Network:
   WSN system includes sensor nodes, sink nodes, and gateway nodes. Data collected by the sensor nodes transmitted to sink node and the data transferred to the cloud server by cloud gateway.

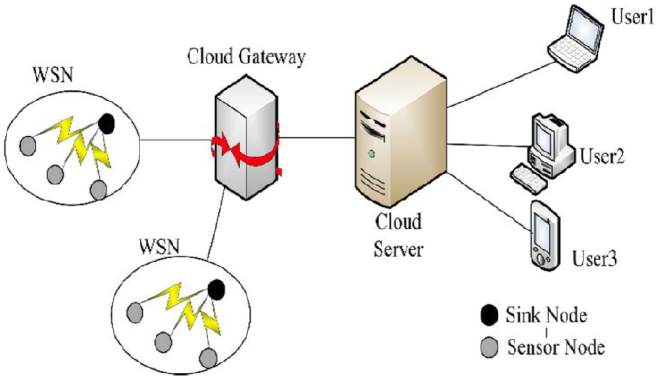

Fig. 3. System model architecture

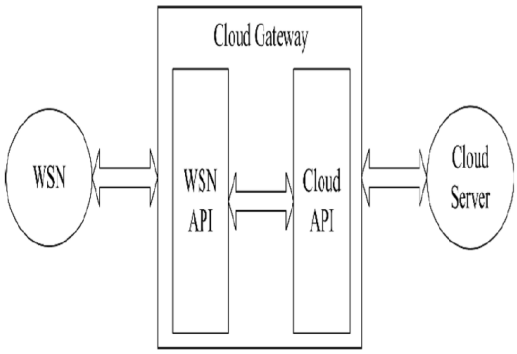

Fig. 4. Cloud gateway

2) Cloud gateway:
Using cloud gateway, we can directly access the cloud applications and sensor devices. WSN brings the sensor data to the gateway node and the gateway node connected to the cloud server and transmits data. The data will be displayed when the user request the cloud server.
The interaction between the gateway and sensor is realized by WSN API and the interaction between the gateway and the cloud apps is realized by cloud API.
3) Cloud server:
Due to limited storage capacity of the sensor nodes, it is impossible for the permanent storage of data. The cloud server could store data collected from wireless sensor network. The cloud storage ensures the safety and reliability of these data.

V. Detection and Prevention of SQL Injection Attack

A. SQL Injection Attacks in Principle

SQL databases are attacked against by direct insertion of code into input variables, which are consisted of the primary form of SQL injection. Some attackers insert malicious code into a string, the server will not respond input values when the input string contains such an SQL statement.

B. SQL Injection Attack

The following example shows how SQL injection attacks realize. SQL injections utilize weakness of a bank's application to misguide the application into running a database backend query or command. Usually, an application of a bank's operation has a menu, which is used for searching customer's personal information, such as the telephone number. The application will execute an SQL query in the database backend.

- SELECT client name, sex,address,date of birth WHERE tel no=123456
  If user enters the string "123456 or 1=1," then the SQL query passes to the database as follows:
  - SELECT client name, sex, address, date of birth WHERE tel no=123456 or 1=1.
  - The condition 1= 1 is always true in database. The query will return all rows in the table, which is not the original intention. The application can be changed so that it accepts one numeric value only. SQL injection attacks can be mitigated by ensuring proper application design, especially in modules that require user input to run database queries or commands.
  - The below script case shows a typical SQL injection. The script creates an SQL query by concatenating hard-coded strings together with an input string. The user is prompted to input the name of a city. If the user enters Seattle, the query is assembled by the script looks like the follows:
- SELECT * FROM Orders Table WHERE Ship City = 'Seattle'
- However, assume that the user enters the following script: SELECT * FROM Orders Table WHERE Ship City = ' Seattle '; drop table Orders Table–'

The semicolon (;) states the end of one query and the start of another query, the double hyphen (–) denotes that the rest of the present line is a comment and should be disregarded. If the edited code is syntactically correct, it will be executed by the server side.

C. SQL Injection Detection Mechanism

Several popular techniques in detecting an attack were being brief as follows:

- Static analysis:
  Static Analysis statically screens web application's source code for vulnerabilities. One of the research conducted used this approach to identify all points on the application code that issues SQL queries to underlying database.
- Dynamic analysis/Runtime monitoring:
  This technique is the compliment of the Static Analysis, which is Dynamic Analysis or Runtime Monitoring. Unlike the static, this analysis locates vulnerabilities of

SQLIA during runtime environment and eliminates the need to modify the web application codes.

- Prepared statement:
  Prepared statement is a fixed query ”template” which is predefined explicitly, providing type specific placeholders for input data.

This method converts vulnerable SQL statements to prepared statements.

D. Reverse Proxy with Mod Security

Reverse proxy for web security is one of the serverside solutions. It can be generalized that most of the common mechanism of a server side solution is to use these proxies as application-firewalls to filter out malicious code or injected request in the case of SQL Injection Attack. Reverse proxies provide a complete separated layer of security for the application level of web applications. They are capable of terminating TCP and SSL protocols besides controlling TCP and SSL handshakes to the clients as well as to the servers.

Several functionalities that make people choose Mod Security configured in a reverse proxy are as follows:

- Real Time Monitoring and Attack Detection:
  Mod Security monitors HTTP traffic in real time to assist attack detection. Similar to other web intrusion detection tool, necessary action is taken prior to the detection of suspicious event.
- Attack prevention and patching:
  Mod Security prevents attack from reaching the web application immediately.
- Flexible rule engine:
  Mod Security allows custom rules to come into action. It makes the common operations simple and complex operation possible through combinations of several rules.
- HTTP traffic logging:
  Mod Security logs each of HTTP session; both requests and responses based on the user’s preferences on the relevant of each data providing a fine granulated filtering.

E. Visualization of Server Logs

Web servers normally are capable of logging traffic in a useful form for marketing analyses, but somehow fail to do so to web applications; especially when it comes to logging the request bodies. That is why most attacks today are performed via POST requests and rendering the systems blind. With HTTP traffic logging, Mod Security has the mechanism to perfectly collect logs for traffics coming in and out from the server.

F. SQL Injection Attack Prevention Methods

When software testers are implementing precautions against malicious input, not only validate user input by testing type, format, length and range, but also consider the architecture and deployment scenarios of their application.

- Validate user Input filtering:

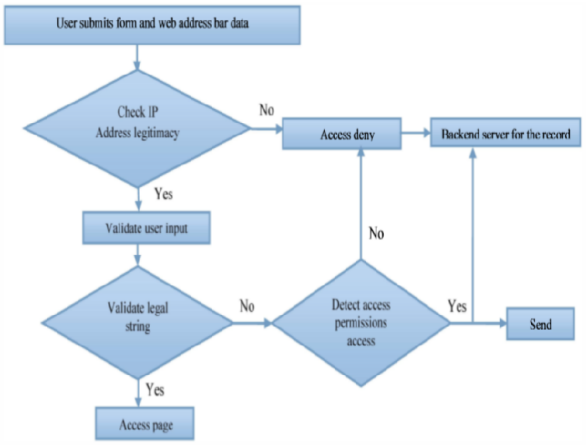

Fig. 5. SQL defence model

Validate user Input filtering input may also helps to protect against SQL injection, such as: “`insertlselectlandl`.

- Use SQL parameters:
  SQL parameters provide length validation and type checking. One benefit of using the Parameters collection is that the input is regarded as a literal value rather than the code to be run. Another benefit is that type and length validation is enforced by client server.
- Use parameterized input:
  Use parameterized input with Stored Procedures Stored procedures are susceptible to SQL injection if they use unfiltered input.

G. SQL Injection Attack Defense Model

This model not only validates input form information, but also detects web address bar information, especially for sensitive character detection. Firstly, server side checks IP address legitimacy. The user is denied to access the login server when the input values are illegal. Secondly, server side checks input values by testing format, length, type, and range. If the input string matches SQL prevention rule, then the user is allowed to access the page. Finally server side verifies the user privilege. If the user’s number exceeds the access permissions, then the user is blocked timely and the system sends a message to system administrator. Server records the injection attack when all verifications are invalid.

VI. Conclusion

In this paper, we proposed proxy server bypassing scheme that bypasses many proxy servers constituting proxy chain. It solves the degradation of quality of web service. Now the most popular proxy server used today is a Web proxy, and it is used to filter contents. Here we covers four fundamental methods and ZigBee network to see how the system works. An additional transparent proxy server is used to re-direct the packets to a special gateway. It is not only possible to connect a sensor network to the internet also realize an interconnection

between different sensor network technologies. And presents a new framework based on cloud gateway. And also highlights SQLIAs and how the combination of a reverse proxy with Mod Security can help organizations to detect and block SQLIA in efficient manner. The SQL injection attack can be prevented by input validation and type-safe SQL parameters methods.

REFERENCES

[1] David Gourley et al., *HTTP: The Definitive Guide*, Sebastopol: O Reilly, Sep 2002.

[2] Ilya Grigoric, *High Performance Browser Networking*, Sebastopol: O Reilly, Mar 2013.

[3] M. Avlesen, S.Spjelkavik, B. Vik, F. T. Johnsen, and T. H. Bloebaum, "Spade: A test framework for SOAP analysis in dynamic environments", in *Proceedings og 9th International Conference on Web Information Systems and Technologies* (WEBIST 2013), 8-10 May, Aachen, Germany, 2013.

[4] P. Bartolomasi, T. Buckman, A. Campbell, J. Grainger, J. Mahaffey, R. Marchand, O. Kruidhof, C. Shawcross, and K. Veum, *NATO network enabled capability feasibility study*, Version 2.0, October 2005.

[5] B. Otgonchimeg, J. T. Kim, S. Y. Lee, and Y. Kwon, "Performance improvement of grid web services based on multi homing transport layer". in *Proceedings of the 6th IEEE International Conference on Computer and Information Technology (CIT06)*, 2006.

[6] Anley C, " Advanced SQL injection SQL server application",[Online] Available: `http://www.creangel.com/papers/advanced sql injection. pdf`.

[7] Liu, A., Yuan, Y., Wijesekera, D., & Stavrou, A., "SQLProb : A Proxy-based Architecture towards Preventing SQL Injection Attacks", in *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 2054 2061, 2009.

[8] Z. Shunyang, X. Du, J. Yongping, and W. Riming, "Realization of Home Remote Control Network Based on ZigBee," in *8th International Conference on Electronic Measurement and Instruments*, (ICEMI07). Guangzhou, China: Chinese Institute of Electronics (CIE) / IEEE August 1618 2007, pp. 344348.

[9] Muthuprasanna, M., Wei, K., & Kothari, S., "Eliminating SQL Injection Attacks - A Transparent Defense Mechanism", in *Eighth IEEE International Symposium on Web Site Evolution WSE06*, pp.22-32, 2006.

[10] K. Gill, S.-H. Yang, F. Yoa, and X. Lu, "A ZigBee-Based Home Automation System, *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 422430, May 2009.

Software Auditing

Clinton Stephen L, Anju Raghunath, Noel P Akkara and Leena Joseph
Vidya Academy of Science & Technology
Thrissur - 680501, India

Sajay K.R
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: sajay.k.r@vidyaacademy.ac.in)

***Abstract*—Software audit is the practice of analysing and observing a piece of software. The word audit is a general term for analysis, and a software audit can consist of several different kinds of review. Software audits involve looking at software for licensing compliance. Others involve looking at how the software works. There are also specific kinds of software audits that show how a software product is configured, and how it is used within a broader IT structure. Software audit involving main feature is implement information security managements.**

Primary objective of the software audit is to identify the vulnerability in web/network based application from internal and external threads. Once the threads are Identified and reported the auditors should also suggest possible remedies. Software auditing is used to check its quality, progress, standards, regulations. The audit is performing to need the various auditing tools. There are three types of tools are using to audit, that is, web based, network based, and vulnerability penetration testing tools. CHECKMARS is the important and most commonly using auditing tool but it is paid tool. Burp-Suite is the freeware tool used for auditing. The software audit is very important for our world. There are 72 organisations in India which are certified for auditing.

***Index Terms*—Software, auditing, Checkmarks, auditing organisations**

I. Introduction

SOFTWARE audit is the practice of analysing and observing a piece of software. The word audit is a general term for analysis, and a software audit can consist of several different kinds of review. Software audits involve looking at software for licensing compliance. Others involve looking at how the software works. There are also specific kinds of software audits that show how a software product is configured, and how it is used within a broader IT structure. Software audit involving main feature is implement information security managements".

Primary objective of the software audit is to identify the vulnerability in web/network-based application from internal and external threads. Once the threads are Identified and reported the auditors should also suggest possible remedies. The audit is performing to need the Various auditing tools. There are 3 types of tools are using to audit. i.e. Web based, network based, and vulnerability penetration testing tools.

The software audit is always called to Software Audit Review. The software audit review is one or more auditors who are not members of the software development organization and conduct "An independent examination of a software product, software process, or set of software processes to assess compliance with specifications, standards, contractual agreements, or other criteria.

Software audits may be performed by in outside firms or consultants. Some audits require teamwork, where a lead auditor may direct the efforts of the team. Software audits will typically rely on specific IT tools that will help with the kind of fact-finding required for the audit in question. That may mean using analysis tools for security or functionality audits.

II. Purpose

Software audits are conducted for the purpose of making sure your business software is properly functioning, meeting standard criteria, and legal, information is secured. If your companys software meets standard criteria, this means that it has been verified that sufficient licenses have been obtained to cover the software that your business is using. Therefore, there is important information gained by conducting a software audit. This information can then ultimately make or break your business. Now people are using different types of software and login to different apps and share personal details(informations), but software are not guaranty to secure our details. The team of hackers are to access our personal data, s through the various hacking software's. The software auditing is not only the review of software but also provide information security management. Information security management of software audit is process of check any vulnerabilities performing the background of the software. The software testing are cannot solve the vulnerabilities of the software and hidden manipulations of the background of the software. Software audit methodology are concentrate the vulnerabilities and hidden patterns of software.The software auditing methodologies are used to find the following activities.

A. Thresholding Method

Thresholding methods are the simplest methods for image segmentation. These methods divide the image pixels with respect to their intensity level. These methods are used over images having lighter objects than background. The selection of these methods can be manual or automatic i.e. can be based

on prior knowledge or information of image features. There are basically three types of thresholding:

1) Security vulnerability
2) Hidden field manipulations
3) Command injection
4) Insecure use of cryptography
5) Server miss-configuration
6) Weak session management
7) Broken links
8) Software leak
9) Logical access control and authorization

III. Types of Software Audit

They are two types of software auditing performing the industry.

A. Internal Audit

Internal software audits are executed by either an in-house team or a team of consultants, whereas external software audits are carried out by external parties. Internal audit reports are used by an organization to check what software is installed across the organization and how the deployed software is utilized across the network. This gives an organization a chance to reassign the unused software, adjust its usage, and manage any internal shortcomings.

B. External Audit

External audits are conducted by vendors or other third parties to ensure that the organizations system is free from under-licensed or pirated software products. If any under-licensed or pirated products are detected, the vendor may demand to resolve the matter immediately, which may result in serious consequences or an unexpected expenditure for the organization. External audit are conducted by certified company's. The certification based on ISO20071.

IV. Lifecycle of Software Audit Process

The software audit process includes the following steps or phases:

A. Planning

Planning is very important step for software auditing. Audit planning is a vital area of the audit primarily conducted at the beginning of audit process to ensure that appropriate attention is devoted to important areas, find potential software problems are promptly identified, software develop phase is completed expeditiously and work is properly coordinated. "Software Audit planning" means developing a general strategy and a detailed approach for the software. The auditor plans to perform the audit in an efficient and timely manner.[Audit planning involves creating an audit plan which enlists all audits to be conducted before release the year, outlines the scope of each audit, and identifies the vulnerabilities of softwares to conduct the audit. Audit preparation involves training personnel and ensuring availability of records and documents for the audit. The planning process auditors are taking the Software Requirement Specification (SRS). The auditors are study the why we create software and find the purpose of software. The internal auditors are tested, and result ton be pass to the external auditors, they are study the all developed phases of the softwares.

- Preliminary assessment and information gathering:
  Although concentrated at the beginning of an audit, planning is an iterative process performed throughout the audit. This is because the results of preliminary assessments provide the basis for determining the extent and type of subsequent testing. If auditors obtain evidence that specific control procedures are ineffective, they may find it necessary to revaluate their earlier conclusions and other planning decisions made based on those conclusions.
- Understanding the software:
  The software auditors are to gather knowledge and purpose of the software. They are detailed study the software by module wise and understand the whole of software. This should include a general understanding of the various business practices and functions relating to the software. The auditor should use this information in identifying potential problems, formulating the objectives and scope of the auditing software.

B. Nature of Software

Understanding the functional requirements of the software system and using hardware. This information provides the auditor to understanding of the risks involved. Though the world is moving towards standardized hardware, differences still exist, and each type of hardware comes with its own vulnerabilities that require specific controls. The auditor should also evaluate the hardware acquisition and maintenance process as a part of his/her preliminary assessment. The auditor needs to understand the type of software used in the organization. The auditor needs to collect details of operating systems, application systems and Database Management Systems used in the organization. The auditor as a part of his preliminary information gathering exercise also needs to collect information relating to network architecture used, the technology to establish connectivity, where firewalls are placed etc. Preliminary assessment of hardware and software would enable planning the audit approach and the resources required for evidence collection.

C. Risk Assessment to Define Audit Objective and Scope

Risk management is an essential requirement of modern software auditing systems where security is important. It can be defined as a process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level. The three security goals of any organization are Confidentiality, Integrity and Availability.
The auditing phase is highly risk more than developing phase, therefore necessary in audit to understand that there is a pay off between the costs and the risks, which are acceptable to the management. For instance, the management might

consciously decide that offsite storage is not required in view of low risks, which are acceptable to the business. In other words it is important to study the management perspective and laid down policy before audit comes to a conclusion of acceptable and unacceptable risks. Therefore, any assessment of the soundness of the software system will necessarily have to study the policies and process of risk management adopted by an organization.

D. Select the Audit Tool

The softwares are created by using different platforms like web-based, network-based VAPT-based and mobile based. The software auditing is analysis and observing the software the software auditing are performed by using various auditing tools. The normal using auditing tools are Check-mars, Nmap, Burp-Suite, Nessus, Rips, OpenVAS, Drib..etc. The auditing tools are different by tool platform.

The tool selection is very important and mandatory phase of software audit lifecycle. So every tool are different platform, web based softwares are audit to use web based auditing tools similar the network-based and vapt based. After planning auditors are understand the software whole configuration and select the appropriate tool for current auditing software.

E. Execution Phase

Next phase of software audit lifecycle. The auditors are selected the tool and after executing the current auditing tool for appropriate auditing software. The tool execution is performing the product code of auditing software are directly execute the tool surface. the tool is check the each software codes like unit wise and module wise. And find the following types of vulnerabilities. If any errors(vulnerabilities) are food the auditing tool they are display to the auditors.

1) Security vulnerability
2) Hidden field manipulations
3) Command injection
4) Insecure use of cryptography
5) Server miss-configuration
6) Weak session management
7) Broken links
8) Software leak

F. Reporting Phase

Reporting phase is another important phase of the software auditing procedure. After the execution of software auditing tools we get the result/report of the software audit. The auditors are analysis the software audit report and find the decisions. its include summary of the software auditing, conclusions, and specifics recommendations RE officially communicated to the clients through a draft report. The reporting phase we get the following result.

1) Executive summary
2) Scope, Methodology and tools used
3) Constraints
4) Summary of results
5) Detailed results of identified vulnerabilities
6) Detailed explanation of the implications of the identified vulnerabilities
7) Business impact and potential risks

G. Issuing Certificate (Final Stage)

After the analysis the records IF current software is without vulnerabilities they are get the SECURITY AUDIT CERTIFICATE . THE certificate are under the ISO27001 global standards. Otherwise they are cannot get security certificate. The auditors are submitted the final report to the audited software company and apply the Third Party Certification. The certificate are get in the 10 days of final report. The Security audit Certified softwares are 100% pure third-party softwares. Certified softwares are safe and ensure the information security.

V. Software Auditing Tools

Software auditing tools are using to analysing and observing the softwares And identified the vulnerability in web/network-based applications. Once the threads are Identified and reported the auditors should also suggest possible remedies. There are many types of tools are using to software audit. Some tools are expensive, and others are free tool for auditing software. The following table are performing the software audit tools. There are 3 types of tools are using to audit, that is:

- Web based
- Network based
- Vulnerability penetration testing tools

A. Checkmarx

Checkmarx is an AST vendor based in Israel with a strong reputation for its SAST solution. Checkmarx has significant presence in North America and Europe, and it also serves the Asia/Pacific (APAC) region. Checkmarx provides CxSAST, which is a SAST product with broad language coverage that provides a variety of options to customize it for specific applications.

Checkmarx's products will appeal to application development and security organizations that are seeking a comprehensive set of AST products and services with a strong set of enterprise-class capabilities and program support services.

- Strengths:
  - Checkmarx offers strong SAST technologies that support a broad variety of programming languages and frameworks, scalability and quick turnaround times via incremental and parallel tests
  - The acquisition and integration of Codebashing enables Checkmarx to deliver innovative training via short, interactive computer-based training models to developers about the vulnerabilities identified in their scans, providing "just in time" training when it's most relevant.

  - Checkmarx gets high marks from users for user experience, ease of use and a generally low learning curve.
- Cautions:
  - Checkmarx does not offer mobile testing such that behavioral testing in a device or emulator.
  - Do not working the java environment. BUT the CxIAST can be used to monitor and analyze data flow of instrumented Java apps during test execution.

B. Veracode

CA Technologies is an AST provider headquartered in the U.S., with a strong presence in the North American market, as well as a presence in the European market. CA Technologies' Veracode offering includes a family of products that provide SAST, DAST and SCA services, as well as IAST (and RASP). Veracode also provides mobile AST. CA Technologies finalized the acquisition of Veracode in April 2017.

Veracode will meet the requirements of organizations looking for complete portfolio of AST services, with broad language and framework coverage and ease of implementation and use.

- Strengths:
  - Easy to use and training is not needed
  - Veracode provides a comprehensive and scalable AST-as-a-cloud service.
  - Veracode offers built-in integration with multiple IDEs, bug-tracking systems and build servers, as well as APIs for integration.
  - Veracode is execute the any type of software's/applications.
- Cautions:
  - Although Veracode has a considerable reputation in the AST space, Gartner inquiries indicate that CA Technologies does not yet have brand recognition as an AST player.
  - Veracode's IAST solution has limited language support and only supports Java. IAST does not support passive testing (as do some IAST competitors) and requires DAST as an inducer.

C. IBM Tools

IBM is a global vendor of IT services and products based in the U.S. IBM provides SAST and DAST desktop tools, including IBM Security AppScan Source, IBM Security AppScan Standard and an enterprise platform (AppScan Enterprise). This includes a centralized management console that enables users to import findings from third-party tools. IBM's cloud services for SAST and DAST (IBM Security Application Security on Cloud). IAST is delivered via the Glassbox agent in AppScan (AppScan Standard, Enterprise and Cloud), which is free to DAST customers, mobile AST (MAST; IBM Mobile Analyzer) and SCA offerings (IBM Security Open Source Analyzer [OSA]).

During the past 12 months, IBM made Open Source Analyzer (OSA) available as a cloud service. IBM improved the Intelligent Code Analysis (ICA) and expanded Intelligent Findings Analytics (IFA) to on-premises customers at no additional cost. Both improve the speed and accuracy of SAST scan results. ICA detects APIs in languages and frameworks and determines the security implications of those APIs to reduce false negatives. IBM IFA uses machine learning to significantly reduce the overall vulnerability count and the number of false positives, and to correlate results and suggest the smallest number of code changes to remediate vulnerabilities.

IBM has a considerable customer base, with an offering combining SAST, DAST and IAST in a single suite of products and services. IBM will appeal to enterprises seeking a single provider of AST technologies, with IBM offerings in adjacent security areas, looking for an AST solution that can provide risk-based management and a full set of enterprise-class capabilities.

- Strengths:
  - IBM has been expanding functionality with an eye toward the needs of DevSecOps. This includes an expanded pallet of language support, splitting the DAST interface into a mode for developers and another for security experts, and running faster, lighter scans for quicker turnaround times.
  - IBM is a large, stable provider of complete AST solutions (SAST, DAST and IAST) and other security products/services with multiregional presence and delivery capabilities.
  - IBM's Application Security Management provides risk-centric, unified reporting and dashboard functionality and the IBM Security Framework and Risk Assessment, the underlying framework to manage business-impacting security risks in applications.
  - IBM is one of the few vendors to allow importation into the reporting dashboard of third-party AST results, such as findings from manual code reviews, penetration testing, vulnerability assessments and competitor AST solutions.
  - IBM Mobile Analyzer offering combines SAST, DAST and IAST for iOS and Android apps, as well as malware analysis.
- Cautions:
  - Some of IBM's newer functionality rests on partnerships, subject to a number of contingencies outside the company's direct control. For example, it was announced that HCL has licensed IBM's AST technology and will build new features. In addition, IBM partners with Prevoty for RASP and WhiteSource for SCA vulnerability and remediation data used by OSA.
  - IBM Mobile Analyzer does not offer behavioral analysis.

D. Qualys

Based in Foster City, California, Qualys is a provider of cloud-based security services with a strong presence in North America and the APAC region, as well as a presence in the European market. Qualys offers Web Application Scanning (WAS), which is a DAST service that is completely automated and integrates with the other Qualys security services in the Qualys Cloud Platform. Qualys provides WAS at an affordable per year subscription, as well as pay-per-scan licensing.

Qualys is does not provide SAST or IAST, and only provides DAST as a cloud service. Organizations looking for a lower-cost, automated DAST service that provides malware scanning should consider Qualys.

- Strengths:
  - Qualys delivers highly scalable, low-cost, largely automated DAST services that will appeal to customers with large-enterprise application portfolios
  - Qualys WAS is quite visible in the DAST market, and WAS is relatively straightforward to deploy and use.
  - Qualys provides extensive, third-party WAF integration and one-click virtual patching with the Qualys WAF.
- Cautions:
  - Qualys doesn't offer IAST, SAST or a dedicated SCA solution, and it has no partnership to offer them.
  - Qualys WAS does not provide certain types of advanced DAST functionality, such as importation of Swagger/OpenAPI specifications to support automated API testing.
  - Qualys WAS does not provide human augmentation options, beyond the Bug Crowd partnership.
  - Qualys mobile AST is limited to dynamically assessing APIs and back-end services, and does not offer behavioral analysis.

E. Rapid7

Rapid7 is a provider of security, data, analytics software and IT services based in Boston, Massachusetts. It has a strong presence in the North American market, as well as the European market. In the AST space, Rapid7 provides DAST as a product and service. Its offering consists of a desktop web app scanner called AppSpider Pro, an on-premises enterprise DAST tool called AppSpider Enterprise and DAST as a service under the name of InsightAppSec. In addition, Rapid7 provides AppSpider Managed Services, which offer the same on-demand DAST in a completely outsourced fashion that also includes vulnerability validation services.

During the past 12 months, Rapid7 launched InsightAppSec and InsightVM on its Insight platform to provide a cloud-based security analytics platform that combines application security data from InsightAppSec, with vulnerability information collected by InsightVM.

Rapid7 should be considered by organizations looking for a competitive alternative to the larger providers for DAST, delivered either as a product, service or fully managed service.

- Strengths:
  - Rapid7 has a strong reputation for comprehensive DAST that can support in-depth manual assessments necessary for custom development use cases, as well as the more automated DAST required to support DevOps.
  - Rapid7 is pursuing a vision of integrated AST and vulnerability management by extending its Insight platform to combine findings from AST with information such as IT log analytics, vulnerability management data and user-behavior analytics gathered from the Insight portfolio.
  - AppSpider has good SDLC and enterprise integration capabilities for a DAST solution, including plug-ins with bug-tracking tools, WAF and IPS products. The vendor recently introduced a Chrome/WebKit integration to support integrated browser functionality with Chrome.
- Cautions:
  - Rapid7 does not provide native SAST capabilities, although it provides SAST through its partnership with Checkmarx.
  - Rapid7 does not support distributed scanning with its DAST offering.
  - Despite industrywide trends toward increased adoption of services, most Rapid7 clients leverage the on-premises implementation, and Rapid7 struggles to be included on shortlists where services are a primary focus.
  - Rapid7 does not provide IAST, nor does it provide behavioral testing. The vendor's mobile AST is limited to analyzing the traffic between the mobile app and the back-end services.

F. Trustwave

Based in Chicago and owned by Singtel since 2015, Trustwave is a worldwide provider of security-related products and services. Trustwave offers a portfolio of application-layer products and services, including web application firewalling, web application vulnerability assessment, network vulnerability scanning and database activity monitoring. Trustwave is a well-known player in the managed security services and Payment Card Industry Data Security Standard (PCI DSS) assessment markets.

During the past 12 months, Trustwave has developed an enhanced DAST scanning engine to better support single-page applications and modern application development. The vendor has also revamped the MST offering to improve the user experience, workflow and testing options.

- Strengths:
  - Trustwave's comprehensive portfolio of technologies and managed security services remains well-known for its support of PCI DSS. Trustwave supports an expansive list of testing and reporting templates tailored to major regulatory requirements. This makes it a good fit for buyers in more-regulated industries.

  - Trustwave provides a number of options for integration in the SDLC, including IDE, bug-tracking, quality testing and several WAF tools, including Trustwave's own WAF and the ModSecurity commercial ruleset.
  - The portfolio of products and service options makes Trustwave suitable for clients with large, varied application portfolios requiring DAST testing.
  - Trustwave client's praise the vendor's support and responsiveness and give high marks for the vendor's flexibility in meeting their requirements.
- Cautions:
  - Trustwave struggles to be included on Gartner client shortlists, where PCI DSS compliance is not a main driver.
  - Trustwave does not offer a SAST product or service, or application vulnerability correlation, nor does it partner to provide them.
  - App Scanner does not provide SCA, nor does it partner for this, although the DAST solution can be used to identify well-known vulnerabilities and misconfigurations in the underlying web and application servers.
  - Mobile AST, delivered via the Managed Security Testing offering, does not include automated static analysis of the code, but it does offer a manual code review service.
  - Trustwave does not offer IAST capabilities, nor does it partner to provide this.

G. WhiteHat Security

Based in the U.S., WhiteHat Security is a global provider of DAST and SAST as a service. WhiteHat has a particularly strong presence in the North American AST services market. WhiteHat's AST suite, Sentinel, provides SAST (with integrated SCA) and DAST as a service, using an on-premises appliance to keep scanning local, when desired, as well as mobile testing delivered via partnership with NowSecure. Sentinel SAST solution can scan both binaries and source code. The results of all of WhiteHat's DAST and SAST scans are reviewed by an expert in WhiteHat's Threat Research Center before delivery to the customer.

- Strengths:
  - Among Gartner clients, WhiteHat Security has a strong reputation as a DAST as-a-service provider.
  - WhiteHat's scalability and continuous testing offering scans web applications for changes and automatically initiates production-safe scanning of those changes. It appeals to organizations with large application portfolios in heavily regulated environments looking to support ongoing vulnerability assessments.
  - WhiteHat's Scout is a fully automated SAST offering integrated into the IDE. It is tuned for high-assurance tests and quick turnaround, which will appeal to clients supporting DevOps use cases. Scout also provides open APIs for custom integrations.
  - WhiteHat's customers continue to value the vendor's strong support services. These include vulnerability verification, manual business logic assessments, and its ability to leverage the vendor's Threat Research Center's engineers to discuss findings and get remediation support.
  - WhiteHat SAST offers an innovative feature called Directed Remediation. This automatically provides custom code patches that can be copied and pasted into the code to fix identified vulnerabilities for a portion of findings (for example, roughly 30% of Java vulnerabilities). WhiteHat will use Attack Vector Intelligence to identify where a single fix can address multiple findings and will combine those findings for reporting and remediation purposes.
- Cautions:
  - WhiteHat's introduction of new features and improvements tends to follow behind other similar features introduced previously by leaders in the market.
  - The introduction of Scout affords improved time to results for SAST findings with Java; however, for languages not supported by Scout, clients in rapid development cycles still express the need for shorter turnaround times to scanning cycles.
  - WhiteHat Security still struggles to compete for inclusion in shortlists, where SAST is the most heavily weighted component. In part, this is due to its limited number of supported languages, relative to other SAST vendors.

VI. Security Auditing Organisations

Software auditing is analysing and observing the piece of software's and find the vulnerability then solve the current issues of software's. But auditors are not members of the software development organization conduct "An independent examination of a software product, software process, or set of software processes to assess compliance with specifications, standards, contractual agreements, or other criteria . The software auditing organisations are needed to certifications by ISO 9000or 9001and ISO 97001. Testing team-members are not granting to audit. The some organisations are got the permission to audit the softwares. Kerala Start-up It Mission (KSUM) is the organization under Kerala State Government is only one government organisation for auditing software's.

In India, there are several auditing organisations in India. The following is an incomplete list of auditing organisations in India.

1) AAA Technologies Pvt Ltd, Mumbai 400072
2) AUDITime Information Systems (India) Ltd, Mumbai 400086
3) AKS Information Technology Services Pvt Ltd, Noida - 201301
4) Aujas Networks Pvt Ltd, Bangalore.

5) AGC Networks, Kurla (West), Mumbai
6) ANB Solutions Pvt. Ltd, Kamla Executive Park, Mumbai
7) BDO India LLP, Maharashtra
8) MS IT Services Pvt Ltd, CMS Haryana
9) Cyber Q Consulting Pvt Ltd, New Delhi
10) Control Case International Pvt Ltd, Andheri(E) Mumbai - 400059
11) Centre for Development of Advance Computing (C-DAC), Keshavagiri, Hyderabad
12) Cyber Security Works Pvt Ltd, Shri M.Ram Swaroop, Anna Salai, Chennai
13) Cigital Asia Pvt Ltd, Bangalore
14) CyberRoot Risk Advisory Pvt Ltd, MG Road, Gurgaon - 122002, Haryana
15) Code Decode Labs Pvt Ltd. Pune
16) Deccan Infotech Pvt Ltd, Bangalore
17) Digital Age Strategies Pvt Ltd, Banglore
18) Deloitte Touche Tohmatsu India Ltd, Worli, Mumbai
19) Ernst & Young Pvt Ltd, Chennai
20) Esec Forte Technologies Pvt Ltd, Gurgaon
21) Finest Minds Infotech Pvt Ltd, Bangalore
22) HCL Comnet Ltd, Noida
23) Haribhakti & Company LLP, Chartered Accountants Andheri
24) isec Services Pvt Ltd, Jogeshwari
25) Indusface Pvt Ltd, Vadodara, Gujarat
26) Imperium Solutions, Maharashtra
27) Kochar Consultants Pvt Ltd, Mumbai 400028.
28) KPMG City, Phase-II, Gurgaon
29) Mirox Cyber Security & Technology Pvt Ltd, Trivandrum, Kerala
30) LTI (A Larsen& Toubro Group Company) L&T House, Ballard Estate, Mumbai
31) Locuz Enterprise Solutions Ltd, Hyderabad
32) Mahindra Special Services Group D4, Saket, New Delhi.
33) Agency for Promotion of Information Technolog, Madhya Pradesh
34) Maverick Quality Advisory Services, Ghaziabad, U.P
35) Lucideus Tech Pvt Ltd, New Delhi
36) Netmagic IT Services Pvt Ltd, Saki Vihar
37) Network Intelligence India Pvt Ltd, Andheri East, Mumbai - 400069
38) Net-Square Solutions Pvt Ltd, Ahmedabad - 380007
39) Paladion Networks JP Nagar, Bangalore
40) PricewaterhouseCoopers Pvt Ltd, Gurgaon
41) Payatu Technologies Pvt Ltd, Bharmal .
42) Panacea InfoSec Pvt Ltd, Delhi
43) Protiviti India Member Pvt Ltd, Gurgaon
44) Qadit Systems & Solutions (P) Ltd, Chennai
45) Qseap InfoTech Pvt Ltd, Navi Mumbai
46) RSM Astute Consulting Pvt Ltd, Mumbai
47) Recon Business Advisory Pvt Ltd, New Delhi
48) Robert Bosch Engineering and Business Solutions Pvt Ltd, Electronic City, Bangalore
49) Sumeru Software Solutions Pvt Ltd, Bangalore 560082
50) Sysman Computers Pvt Ltd, Mumbai
51) SISA Information Security Pvt Ltd, Bangalore
52) STQC Directorate, New Delhi
53) Suma Soft Pvt Ltd, Pune, Maharashtra
54) Security Brigade InfoSec Pvt Ltd, Mumbai, Maharashtra
55) Sify Technologies Ltd Taramani, Chennai
56) Sandrock eSecurities Pvt Ltd, Delhi
57) SecurEyes Techno Services
58) SecureLayer7 Technologies Pvt Ltd, Pune
59) Sonata Software Ltd, Bangalore - 560019.
60) TAC InfoSec Private Limited, Mohali-160055
61) Talakunchi Networks Pvt Ltd, Goregaon West, Mumbai - 400104
62) TATA Communications Ltd, Mumbai - 400098
63) Torrid Networks Pvt Ltd, Noida
64) TCG Digital Solutions Pvt Ltd, Kolkata
65) Tech Mahindra Ltd, Pune - 411004
66) Trusted Info Systems Pvt Ltd, New Delhi - 110048
67) TV SD South Asia Pvt Ltd, Mumbai - 400 072
68) ValueMentor Consulting LLP, Koratty, Kerala - 680308
69) Varutra Consulting Pvt Ltd, Corporate Office Maharashtra
70) Vista Infosec Pvt Ltd, Mumbai
71) Wipro Ltd, Gurgaon, Haryana
72) Xiarch Solutions Pvt Ltd, New Delhi

VII. Advantage of Software Audit

1) *Audit helps to detect and prevent errors and frauds:* Software auditing main duty is to detect errors and frauds, preventing such errors and frauds and taking care to avoid such frauds. Thus, even though all organisations do not have compulsion to audit.
2) *Preventing information loss:* Can you imagine your crucial software data is hacked and its with your competitor or any unwanted hands? Sensitive information of users if more important, and it should be highly secured.
3) *Speed and accuracy:* The important advantage of the software audit is speed and accuracy. Clearly, if an auditors can use run source code of the software through auditing tools, then the job can be done very quickly and get the accurate result (vulnerabilities or not) of the software.
4) *Models:* The more complex auditing tools can provide another level of the aid by generating computer models and simulations for auditor. This is very useful when an auditor is working to study the software design approach and how it can be restructured. The models are used to judge the potential for risk(from mistakes, fraud and other problems) in current system.
5) *Essential part of compliance standards or certifications for your software:* Software audit helps shape information security strategy through identifying vulnerabilities and quantifying their impact and likelihood so that they can be managed proactively; budget can be allocated, and corrective measures implemented.

6) *Maintaining the reputation of the software:*
   Over the audit of software completion and get the audit certificate, the reputation of the company is enhanced in the meanwhile ensuring the growth of the organisation.
7) *Effectiveness:*
   It deals with informations of the software being relevant and perstinent to the business process as well as being delivered in a timely, correct, consistent and usable manner. Protect the information security management system.
8) *Confidentiality:*
   Software audit concerns protection of sensitive information from unauthorized disclosure.
9) *Transparency:* One of the main advantages of software audit is transparency. Transparency means that the auditors are done the phases of software auditing is viewable to the software developers.

VIII. Disadvantages of software audit

1) *Strengths:*
   On the downside, an assisted audit bases much of its usefulness on the auditing software. If the software is out date or is not designed to be used for an auditing purpose, then it can do more damage than good, requiring frequent updates and customization.
2) *Training issues:*
   The software auditing is performed by the certified and experienced auditors. sometimes auditing is performed by new auditors they are not get the training, then they are cant know the how to use auditing tools correctly, which can lead the bad decisions and loss the high costs in time and money for audited firms.
3) *Problems in remedial measures:*
   After execution of auditing tools and we get some issues and auditors are take the remedial for solving the issues of software. But take the remedials are act the current software. It acts the some issues for configuration of the software.
4) *Expensive:*
   The auditors are auditing the software using various tools but some more tools are costly (paid) tools. So Initial costs of tools are high. The paid tools are update every year then updates are getting to pay the money.

IX. Conclusion

Software audit is the practice of analysing and observing a piece of software. Software audits involve looking at software for licensing compliance. Others involve looking at how the software works. There are also specific kinds of software audits that show how a software product is configured, and how it is used within a broader IT structure. Software auditing is used to check its quality, progress, standards, regulations.

The software audit is process of identify the vulnerability in web/network-based application from internal and external threads. Once the threads are Identified and reported the auditors should also suggest possible remedies. Software audit are performing the use auditing tools. CHECKMARX is the important and most commonly using auditing tool but it is paid tool. Burp-Suite is the most commonly using free tool for auditing. The software audit is very important for our world.

References

[1] "Software Audit", in *Techopedia* [Online]. Available: https://www.techopedia.com/definition/9444/software-audit.
[2] Indian Computer Emergency Response Team, "Empanelled Information Security Auditing Organisations by Cert-IN". [Online] Available: https://www.cert-in.org.in/PDF/Empanel_org.pdf .

Comparison of Encryption Algorithms in Cloud Environment

Deepak K, Sreeshma K S and Aneesha T A
Vidya Academy of Science & Technology
Thrissur - 680501, India

Siji K B
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: siji.k.b@vidyaacademy.ac.in)

***Abstract*—With the tremendous growth of sensitive information on cloud, cloud security is getting more important than even before. The cloud data and services reside in massively scalable data centers and can be accessed everywhere. The growth of the cloud users has unfortunately been accompanied with a growth in malicious activity in the cloud. More and more vulnerabilities are discovered, and nearly every day, new security advisories are published. Millions of users are surfing the Cloud for various purposes; therefore, they need highly safe and persistent services. The future of cloud, especially in expanding the range of applications, involves a much deeper degree of privacy, and authentication. We made comparisons among DES, AES, and RSA algorithms to find combinatorial optimization solutions, which fit Cloud environment well for making cloud data secure and not to be hacked by attackers.**

***Index Terms*—Cryptography, Encryption, Decryption, DES, AES, RSA.**

I. Introduction

CLOUD computing is emerging as a key computing platform for sharing resources that include infrastructure, software, applications, and business processes. Gartner predicts by 2015, 10% of overall IT security enterprise capabilities will be delivered in the cloud, with focus on messaging, web security and remote vulnerability assessment. Other focus areas will include data-loss prevention, encryption, and authentication, as technologies aimed to support cloud computing mature. The notion behind cloud computing is that work done on the client side can be moved to some unseen cluster of resources over the internet. Cloud Service Provider (CSP) maintains database and applications for the users on a remote server and provide independence of accessing them from any place through a network. There are three major cloud service categories: software-as-a-service (SaaS), platform-as-a-service (PaaS) and infrastructure-as-a-service (IaaS).

Information is an asset that has a value like any other asset. As an asset, information needs to be secured from attacks. Now-a-days security becomes an essential feature in almost all area of communication. While sending a message to a person over an insecure channel such as internet we must provide confidentiality, integrity, authenticity and non-repudiation. These are the four major security aspects or goals.

The process of encoding the plaintext into cipher text is called Encryption and reverse the process of decoding ciphers text to plaintext is called Decryption. This can be done by two techniques symmetric-key cryptography and asymmetric key cryptography. Symmetric key cryptography involves the usage of the same key for encryption and decryption. But the Asymmetric key cryptography involves the usage of one key for encryption and another, different key for decryption. Secret key cryptography includes DES, AES algorithms etc. and public key cryptography includes RSA algorithm. We compare and analyzed algorithms DES, AES and RSA algorithms

II. Cryptography

Cryptography is derived from Greek language cryptos means hidden and grafos meaning write or speak which means study of hiding information. It is the science of securing data. Cryptography is a science of using mathematics to encrypt and decrypt data. Cryptography enables to store important data or transmit it across insecure networks so that it cannot be read by anyone except the intended recipient. Cryptography examples include the security of the ATM cards, computer passwords and electronic commerce which all depend upon cryptography.

A. Purpose of Cryptography

Cryptography is necessary when communicating over any medium such as internet. Mostly used for communicating over un-trusted medium. To send information over an un-trusted medium there are some specific requirements such as Authentication: Authentication is a process of identifying an individual, such as based on username and password. Privacy: Privacy is ensuring the sender that the message can be read by the intended receiver and no one else. Integrity: Assuring the receiver that the received message has to been altered in any way from the original. Non-repudiation: It is a method of guaranteeing message transmission between two parties. Successful completion of message sent and received.

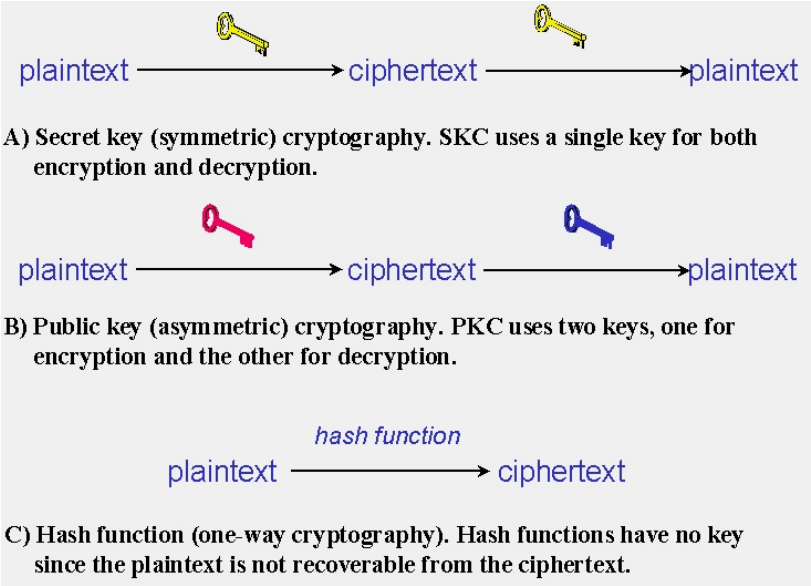

Fig. 1. Two types of cryptographic algorithms

III. Methods of Encryption

There are several blocks in an encryption method, the two main blocks are the algorithms and the key. Algorithms are the complex mathematical formulas that dictate the rules of how the plaintext will be converted into cipher text. Key is a set of random bits that will be inserted into the algorithm. Two users can communicate via encryption, they must use the same algorithm and the same key. In some encryption cases, the receiver and sender use the same key and in other encryption cases they must use different keys for encryption and decryption process. There are three different methods of encryption namely symmetric, asymmetric and hash function method for encryption and decryption.

Cryptography algorithms use either symmetric keys or asymmetric keys. Symmetric keys are also called secret keys which uses a single key for encryption and decryption. Asymmetric keys are also called as public keys which makes use of two different keys for encryption and decryption.

A. Symmetric Cryptography

In symmetric cryptography both the parties i.e., the sender and the receiver will be using the same key for encryption and decryption process as show in Fig 1 (A).

A sender uses a key to encrypt plain text into cipher text and sends it to the receiver. Receiver uses the same key to decrypt the cipher text into plain text. As told above symmetric key is also called as secret key, because this type of encryption depends on each user, there can be more than one user but the users should keep the key a secret and properly protect it. If this key goes into the intruders hand, that intruder has the access to decrypt any intercepted message encrypted with this key.

The following are examples of symmetric key algorithms:

- DES
- AES

The following list outlines the advantages and disadvantages of symmetric key systems:

1) Advantages:
   - Much faster than asymmetric method.
   - Hard to break the key if large key size is used.
   - Compared to asymmetric systems, symmetric algorithms scream in speed.
2) Disadvantages:
   - Key distribution: The key must be delivered in a proper way.
   - Scalability
   - Limited security

B. Asymmetric cryptography

In asymmetric cryptography two different keys are used for encryption and decryption. In this type of cryptography, a pair of keys is made up of one public key which can be known to everyone and one private key which is known only to the owner. Example of asymmetric cryptography is demonstrated for better understanding. In asymmetric cryptography two different keys are used for encryption and decryption. In this type of cryptography, a pair of keys is made up of one public key which can be known to everyone and one private key which is known only to the owner. Example of asymmetric cryptography is demonstrated for better understanding.

The following are examples of asymmetric key algorithms:

1) RSA
2) Elliptic Curve Cryptosystem (ECC)
3) Diffie-Hellman
4) Digital Signature Standard (DSS)

IV. Detailed Description of Common Encryption Algorithms

The generation, modification and transportation of keys have been done by the encryption algorithm. It is also named as cryptographic algorithm. There are many cryptographic algorithms available in the market to encrypt the data. The strength of encryption algorithm heavily relies on the computer system used for the generation of keys. Some important encryption algorithms are discussed here:

A. Data Encryption Standard (DES)

DES is one of the most widely accepted, publicly available cryptographic systems. It was developed by IBM in the 1970s but was later adopted by the National Institute of Standards and Technology (NIST), as Federal Information Processing Standard 46 (FIPS PUB 46). The Data Encryption Standard (DES) is a block Cipher which is designed to encrypt and decrypt blocks of data consisting of 64 bits by using a 64-bit key.

The Data Encryption standard is used to protect electronic data. DES algorithm uses symmetric block cipher for encrypting and decrypting data. Encryption converts data into gibberish language called cipher text. Decrypting the cipher text gives us back the original data that is plaintext. Converting the information from cipher to plain we use a standard form of algorithm called Symmetric algorithm.
DES takes an input of 64bits and the output is also of the same size. The process requires a second input, which is a secret key with length of 64bits. Block cipher algorithm is used where message is divided into blocks of bits. Block cipher is used for encryption and decryption.DES works on 64 bits of data at a time. Each 64 bits of data is iterated on from 1 to 16 times (16 is the DES standard). For each iteration a 48-bit subset of the 56-bit key is fed into the encryption block represented by the dashed rectangle above. Decryption is the inverse of the encryption process. There are many attacks and methods recorded till now those exploit the weaknesses of DES, which made it an insecure block cipher. Despite the growing concerns about its vulnerability, DES is still widely used by financial services and other industries worldwide to protect sensitive on-line applications.

1) Advantages:
   - By using DES, input message of 64 bits can be encrypted using the secret key length of 64 bits.
   - The encrypted key is cipher key which is expanded into a larger key, which is later used for other operations.
   - DES is hard to attack.
   - DES is very hard to crack because of the number of rounds used in encrypting message.
   - DES is faster when compared RSA Encryption Algorithm.
   - DES has high level of security. It is completely specified and very easy to understand. It is adaptable to different applications. Data rates are high. DES can be validated and Exportable.
2) DES Application:
   - DES algorithm was made mandatory for all financial transactions by the U.S government which involves electronic fund transfer.
   - High speed in ATM.
   - It is used for secure video teleconferencing.
   - Used in Routers and Remote Access Servers
   - It can be used by federal departments and agencies when they require cryptographic protection for sensitive information.

B. Advanced Encryption Standard (AES)

AES is the new encryption standard recommended by NIST to replace DES in 2001. AES algorithm can support any combination of data (128 bits) and key length of 128, 192, and 256 bits. The algorithm is referred to as AES-128, AES-192, or AES-256, depending on the key length. During encryption decryption process, AES system goes through 10 rounds for I28-bit keys, 12 rounds for I92-bit keys, and 14 rounds for 256-bit keys in order to deliver final cipher-text or to retrieve the original plain-text. AES allows a 128-bit data length that can be divided into four basic operational blocks. These blocks are treated as array of bytes and organized as a matrix of the order of 4x4 that is called the state.For both encryption and decryption, the cipher begins with an AddRoundKey stage. However, before reaching the final round, this output goes through nine main rounds, during each of those rounds four transformations are performed:

1) Sub-bytes
2) Shift rows
3) Mix-columns
4) Add round Key
5) In the final (10th) round, there is no Mix-column transformation.

Decryption is the reverse process of encryption and using inverse functions: Inverse Substitute Bytes, Inverse Shift Rows and Inverse Mix Columns. Each round of AES is governed by the following transformations:

1) Substitute Byte transformation: AES contains 128-bit data block, which means each of the data blocks has 16 bytes. In sub-byte transformation, each byte (8-bit) of a data block is transformed into another block using an 8-bit substitution box which is known as Rijndael Sbox.
2) Shift Rows transformation: It is a simple byte transposition, the bytes in the last three rows of the state, depending upon the row location, are cyclically shifted. For 2nd row, 1-byte circular left shift is performed. For the 3rd and 4th row 2-byte and 3-byte left circular left shifts are performed respectively.
3) Mix columns transformation: This round is equivalent to a matrix multiplication of each Column of the states. A fix matrix is multiplied to each column vector. In this operation the bytes are taken as polynomials rather than numbers.

4) Addroundkey transformation: It is a bitwise XOR between the 128 bits of present state and 128 bits of the round key. This transformation is its own inverse.
5) In the final (10th) round, there is no Mix-column transformation.

C. Rivest-Shamir-Adleman (RSA)

RSA is designed by Ron Rivest, Adi Shamir, and Leonard Adleman in 1978. It is one of the best-known public key cryptosystems for key exchange or digital signatures or encryption of blocks of data. RSA uses a variable size encryption block and a variable size key. It is an asymmetric (public key) cryptosystem based on number theory, which is a block cipher system. It uses two prime numbers to generate the public and private keys. These two different keys are used for encryption and decryption purpose. Sender encrypts the message using Receiver public key and when the message gets transmit to receiver, then receiver can decrypt it using his own private key. RSA operations can be decomposed in three broad steps; key generation, encryption and decryption. RSA have many flaws in its design therefore not preferred for the commercial use. When the small values of p and q are selected for the designing of key then the encryption process becomes too weak and one can be able to decrypt the data by using random probability theory and side channel attacks. On the other hand, if large p and q lengths are selected then it consumes more time and the performance gets degraded in comparison with DES. Further, the algorithm also requires of similar lengths for p and q, practically this is very tough conditions to satisfy. Padding techniques are required in such cases increases the systems overheads by taking more processing time.

1) Key Generation Procedure:
   a) Choose two distinct large random prime numbers p and q such that $p < q$.
   b) Compute $n = pq$.
   c) Calculate: $\phi(n) = (p-1)(q-1)$.
   d) Choose an integer e such that $1 < e < \phi(n)$.
   e) Compute d to satisfy the congruence relation $de = 1 \pmod{\phi(n)}$; d is kept as private key exponent.
   f) The public key is (n, e) and the private key is (n, d). Keep all the values d, p, q and $\phi(n)$ secret.
2) Encryption:
   a) Plaintext: $P < n$.
   b) Ciphertext: $C = P^e \pmod{n}$.
3) Decryption:
   a) Ciphertext: C
   b) Plaintext: $P = C^d \pmod{n}$.

V. Comparisons of Encryption Algorithms

In the table below a comparative study between AES, DES and RSA is presented in to eighteen factors, which are Key Size, Block Size, Ciphering & Deciphering key, Scalability, Algorithm, Encryption, Decryption, Power Consumption, Security, Deposit of keys, Inherent Vulnerabilities, Key used, Rounds, Stimulation Speed, Trojan Horse, Hardware & Software Implementation and Ciphering& Deciphering Algorithm. DES and AES belong to symmetric algorithm, which characted by high speed and efficiency. DES is the first detail open encryption algorithm. AES has been designed as new standard by National Institute of standards and technology (NIST) of United States in Oct 2000. AES has the same reliability with triple DES at least, but much faster in execute. With key expansion and round key for each encryption round, AES has the ability of resisting key exhaustion attack. With shorter execute time, AES is fit for encryption/decryption of large amount of data. Beside of all, as symmetric algorithm, AES has a fatal security concern on encryption/ decryption key transfer and changing, which is easy to be captured. Its difficult to safely manage huge amount of keys when communicating with large-scale of costumers, which limits the usage in cloud environment. To store the keys, a physical key management server can be deployed in the users premises. This encryption solution protects data and keys and guarantees them remain under users control and will never be exposed in storage or in transfer.

TABLE I
Comparisons of DES, AES and RSA of Encryption and Decryption Times

S.NO	*Algor*	*Pack Size (KB)*	*Encrypt Time (Sec)*	*Decrypt Time (Sec)*	*Buff Size*
1	DES	153	3.0	1	157
	AES		1.6	1.1	152
	RSA		7.3	4.9	222
2	DES	118	3.2	1.2	121
	AES		1.7	1.2	110
	RSA		10.0	5.0	188
3	DES	196	2.0	1.4	201
	AES		1.7	1.24	200
	RSA		8.5	5.9	257
4	DES	868	4.0	1.8	888
	AES		2.0	1.2	889
	RSA		8.2	5.1	934
5	**DES**	312	3.0	1.6	319
	AES		1.8	1.3	300
	RSA		7.8	5.1	416

RSA consumes the largest memory size and encryption time. The security of RSA is based on the complex of prime decomposition, so RSA needs a prime number big enough to produce a long key. RSA is usually fit for encryption/decryption of small data. The key of RSA algorithm used to ensure the security of data in Cloud Computing is usually 1024-bit. To integrate the advantages of different algorithms, a combination solution was presented. In the solution, RSA

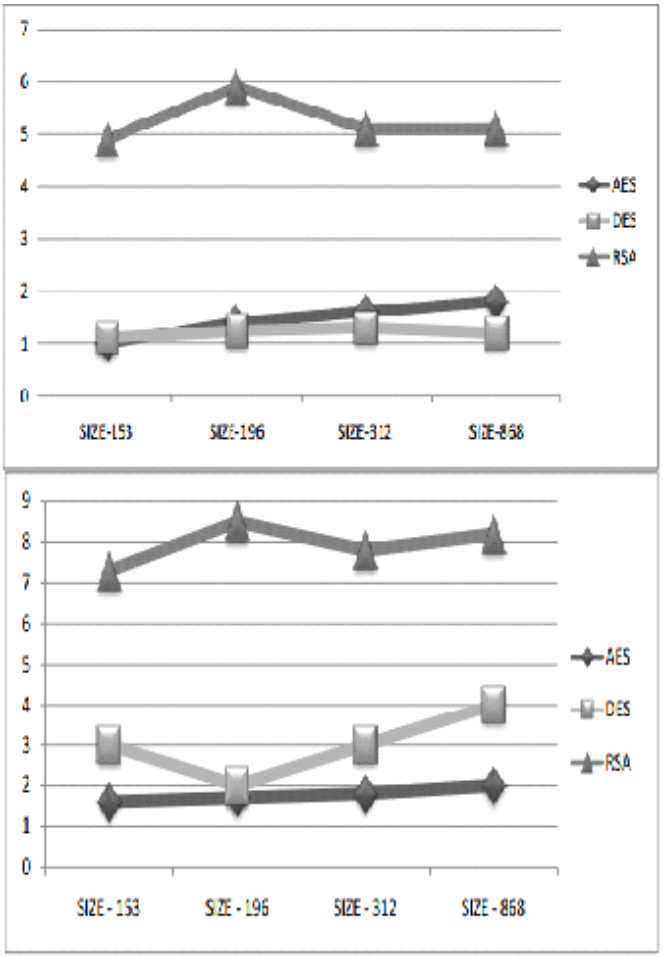

Fig. 2. Comparison of encryption and decryption times

was employed to authentication field and session key, which only has small data 1110 Industrial Engineering, Computation and Information Technologies amount. AES was employed to encryption/decryption of large amount of plaintext data. The execute time was near AES. Meanwhile, AES key encrypted by RSA was transferred between Cloud Service Provider (CSP) and customer, which avoid the risk in transfer process and with no need to deploy an additional physical key management server. The four text files of different sizes are used to conduct four experiments, where a comparison of three algorithms AES, DES and RSA is performed.

1) Evaluation parameters:
   Performance of encryption algorithm is evaluated considering the following parameters.
   a) Encryption time
   b) Decryption time

The encryption time is considered the time that an encryption algorithm takes to produces a cipher text from a plain text. Encryption time is used to calculate the throughput of an encryption scheme, is calculated as the total plaintext in bytes encrypted divided by the encryption time. Comparisons analyses of the results of the selected different encryption scheme are performed.

VI. Conclusion

Encryption algorithm plays very important role in communication security. Our research work surveyed the performance of existing encryption techniques like AES, DES and RSA algorithms. Based on the text files used and the experimental result it was concluded that AES algorithm consumes least encryption and RSA consume longest encryption time. We also observed that Decryption of AES algorithm is better than other algorithms. From the simulation result, we evaluated that AES algorithm is much better than DES and RSA algorithm. Our future work will focus on compared and analyzed existing cryptographic algorithm like AES, DES and RSA. It will include experiments on image and audio data and focus will be to improve encryption time and decryption time.

References

[1] William Stallings, *Cryptography and Network security: Principles and Practices*, Prentice Hall Inc., second edition, 1999.
[2] Paul C. van Oorschot Alfred J. Menezes and Scott A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.
[3] Behrouz A. Forouzan, *Cryptography and Network Security*, Tata McGraw-Hill, 2007.
[4] William J Caelli, Edward P Dawson, and Scott A Rea. "PKI, elliptic curve cryptography, and digital signatures", Computers and Security, 18 (1999) 47-66
[5] Lawrence C. Washington, *Elliptic Curves: Number Theory and Cryptography*, CRC Press, 2003.
[6] Surya A, Efendi R, Sutikno S, "An implementation of El Gamal and elliptic curves cryptosystems",1998 IEEE Asia-Pacific Conference on Circuits and Systems, pages 483.
[7] Uma Somani, Kanika Lakhani and Manish Mundra, "Implementing Digital Signatures with RSA Encryption: Data Security of Cloud in Cloud Computing", 1st International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 211-216, 2010.
[8] Phil Zimmermann, *An Introduction to Cryptography*, Network Associates, Inc, 2000.
[9] Ajay Kakkar, M. L. Singh and P.K. Bansal, "Comparison of Various Encryption Algorithms and Techniques for Secured Data Communication in Multinode Network", *International Journal of Engineering and Technology*, Volume 2 No. 1, pp. 87-92, January 2012.
[10] Aman Kumar, Sudesh Jakhar and Sunil Makkar, "Comparative Analysis between DES and RSA Algorithms", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume2, Issue 7, pp. 386-391, July 2012.

Cloud Computing Storage, Simulation Tools and Security: A Survey

Divya K M, Jahana Shirin Jafar, Riya Antony and Soniya Varghese
Vidya Academy of Science & Technology
Thrissur - 680501, India

Reji C Joy
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: reji.c.j@vidyaacademy.ac.in)

***Abstract*—Cloud computing has become an emerging technology infrastructure in the IT industry. Cloud computing is a service over a network connection which shares enormous amount of resources rather than having to build the infrastructure in house. Cloud computing are often outlined as a computing surroundings wherever computing wants by one party are often outsourced to a different party and once would like be arise to use the computing power or resources like information or emails, they will access them via web. This paper is for anyone who will have recently detected regarding cloud computing and desires to grasp a lot of regarding cloud computing. During this paper, we described cloud computing, architecture of cloud computing, storage of cloud computing, and different simulation tools and security of cloud computing.**

***Index Terms*—Cloud, security, tools**

I. Introduction

Cloud computing is the use of various services, such as software development platforms, servers, storage and software, over the internet, often referred to as the ”cloud”. Cloud Computing provides a surroundings for resource sharing in terms of frameworks, middlewares and application development platforms, and business applications. Cloud computing system can be divided into two sections: the front end and the back end.Front is what the consumer (user) sees ,rear end is that the cloud of the system. For it to be considered “cloud computing,” you need to access your data or your programs over the Internet, or at the very least, have that data synced with other information over the Web. In a big business, you may know all there is to know about what’s on the other side of the connection; as an individual user, you may never have any idea what kind of massive data processing is happening on the other end. The end result is the same: with an online connection, cloud computing can be done anywhere, anytime.

II. Some Basic Ideas of Cloud Computing

A. Architectural Layers of Cloud Computing

1) The hardware layer:
This layer deals with the physical assets of the cloud.That including routers, servers, switches, cooling systems and power.
2) The infrastructure layer:
It is also called as virtualization layer. The infrastructure layer makes a pool of storage capacity and computing resources.
3) The platform layer:
The platform layer based on top of the infrastructure layer, and this layer comprises of operating systems and requisition structures
4) The application layer:
The application layer includes actual cloud provisions. e.g. Business Applications, Multimedia & Web Services. [2]

Fig. 1. Cloud computing environment

B. Service Models of Cloud Computing

1) Infrastructure as a Service (IAAS):
IaaS provides IT infrastructures (processing, storage, networks, and other fundamental computing resources).

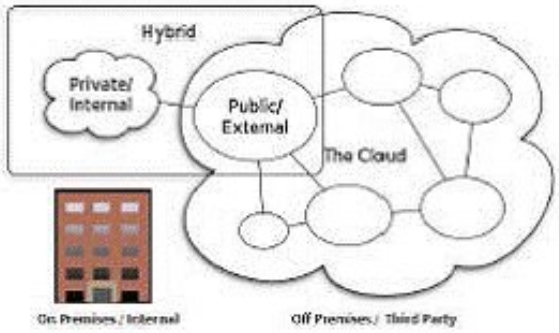

Fig. 2. Cloud computing deployment model

IaaS cloud provides Virtualization in order to integrate/decompose physical resources in an ad-hoc manner to meet growing or shrinking resource demand from cloud consumers. Ex: Amazon's EC2.[2]

2) Platform as a Service (PAAS):
PaaS provides a development platform that supports the full "Software Lifecycle" which allows cloud consumers to develop their cloud services and applications. PaaS offers a development platform that hosts both completed and in-progress cloud applications. Ex:Google AppEngine.[1]
3) Software as a Service (SAAS):
SaaS is a model for the distribution of software where customers access software over the Internet. In SaaS, a service provider hosts the application at its data center and a customer accesses it via a standard web browser. SaaS only hosts completed cloud applications.[2]

C. Deployment of Cloud Computing

1) Public Cloud:
The public cloud is defined as computing services offered by third-party providers over the public Internet, making them available to anyone who wants to use or purchase them. They may be free or sold on-demand, allowing customers to pay only per usage for the CPU cycles, storage or bandwidth they consume.[2]
2) Private Cloud:
Private cloud refers to a model of cloud computing where IT services are provisioned over private IT infrastructure for the dedicated use of a single organization. A private cloud is usually managed via internal resources. [3]
3) Community Cloud:
A community cloud is a collaborative effort made for sharing infrastructure between multiple organizations. It forms into a degree of economic scalability and democratic equilibrium. The community cloud is managed and secured by all the participating organization or by a third party service provider.[1]
4) Hybrid Cloud:
Hybrid cloud is a combination of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds) [1]

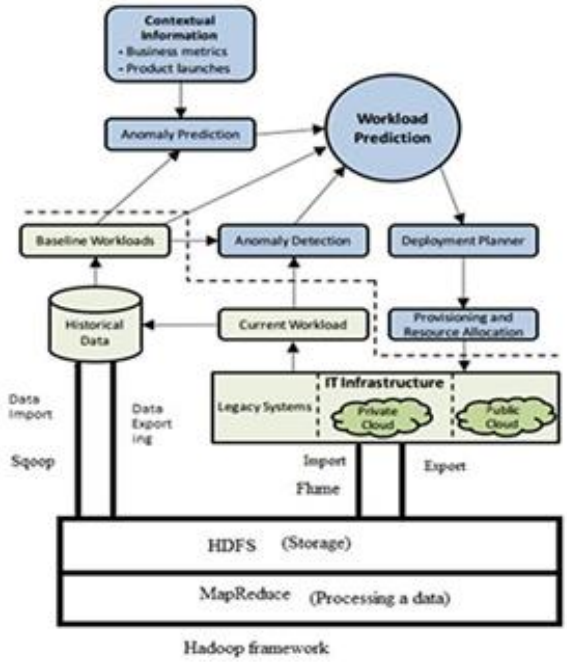

Fig. 3. Architecture for importing and exporting of data with help of Sqoop and flume

III. Cloud Storage

1) Anomaly Prediction:
Alternative sources of statistics may venture Different degrees of planning, and they can be available in different formats.
2) Anomaly Detection:
Anomaly detection is an effective means of identifying unusual or unexpected events and measurements within a web application environment. second line of guard against loss of execution brought on by odd workloads. second line of guard against loss of execution brought on by odd workloads. Operation of this module depends on the workload saw in a given time and standard workloads.
3) Workload Prediction:
The Workload Prediction module completes the interpretation of watched or sudden difference in estimations to the business effect of conceivable interruptions.
4) Deployment Planning:
The Deployment Planning component of our framework is responsible for advising actionable steps related to deployment of resources in a cloud infrastructure to react to failures or anomalies in the system. Automation engine in the Provisioning and Resource Allocation module of the system executes these steps. The execution of cloud workloads during the provisioning of resources is

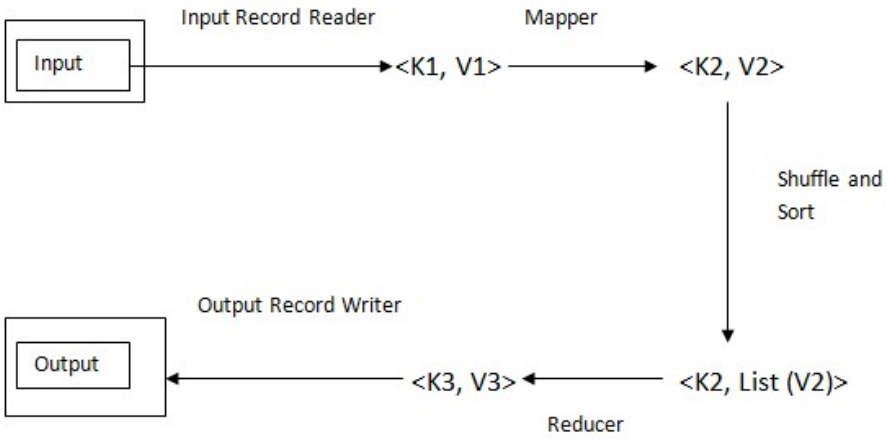

Fig. 4. Map Reduce process

a challenging task. This is because there is a period of waiting time between the moment resources are requested and the provision of resources by the cloud providers and the time, they are actually available for workload execution.

5) Provisioning and Resource Allocation:
Acknowledgment of the arranging choice performed by the Deployment Planner module.
6) Historical Data:
Historical database maintains old data. historical database interacts to Hadoop frame work and in-between these two Sqoop is useful for import and export the data from database to Hadoop and flume is useful for loading the data from enhanced cloud to Hadoop. Hadoop have a two core components:
   - HDFS
   - Map Reduce
7) HDFS:
HDFS means Hadoop Distributed File System. It is useful for storing the data in the form of blocks. HDFS have the following process:
   - Name Node: Storing the Metadata (data about data).
   - Data Node: Storing the actual or original data.
   - Secondary Name Node: It stores the back of Name Node Metadata.

   Challenges of HDFS:
   - Low-latency data access is not there.
   - Arbitrary modifications are allowed.
   - Lots of small files are an issue.
8) Map Reduce:
Map Reduce is useful for processing the data. It is mainly having a `map()` and `Reduce()` functions.
Map Reduce Processes consist of the following:
   - Job Tracker: Assign the job tasks to the task tracker. And all so allot the Job ids.
   - Task Tracker: Executes the job tasks and gives back to job tracker and again JT send report to the JT.

Fig 4 shows the details of the MapReduce processes.
The Fig 4 shows the how to Record reader read the input data (it may be image or video or text) and it will convert into Key and value (`<K1 (line offset, V1 (line content)>`) and these values take the `maper()` method. The method converts into `K2, V2` and these are passed to shuffle and shot. After that it converts in to `K2, list(v2)`. Now reducer takes those one and reduce the redundant values not for keys and converts to `<K3, V3>`. Finally record writer convert into output.
Cluster setup in Hadoop:
Hadoop supports the parallel distributed processing. So, here adding the nodes parallel in cluster. Adding the nodes is a called a commissioning and deleting the nodes from cluster is called decommissioning.

IV. Cloud Simulation Tools

The conceptual cost for buying the services of different services providers may lead to increase in budget or wastage of money and time. So the solution to this problem is trying out the simulation tools. these tools may include the different algorithms used by different service providers. The use of simulation tools leads to decrease in overall conceptual or operational cost of the organizations. There are different simulation tools available in the market. Simulation tools used for the purpose of simulation and modeling.

A. Some Common Tools

There are various cloud simulation tools available today. Some of them are explained as follows:

1) Cloudsim:
This simulation tool used in largedata centers. The CloudSim toolkit supports both system and behavior modeling of cloud system components such as data centers,virtual machines.
The CloudSim simulator is a layered architecture.
   a) Network Layer:

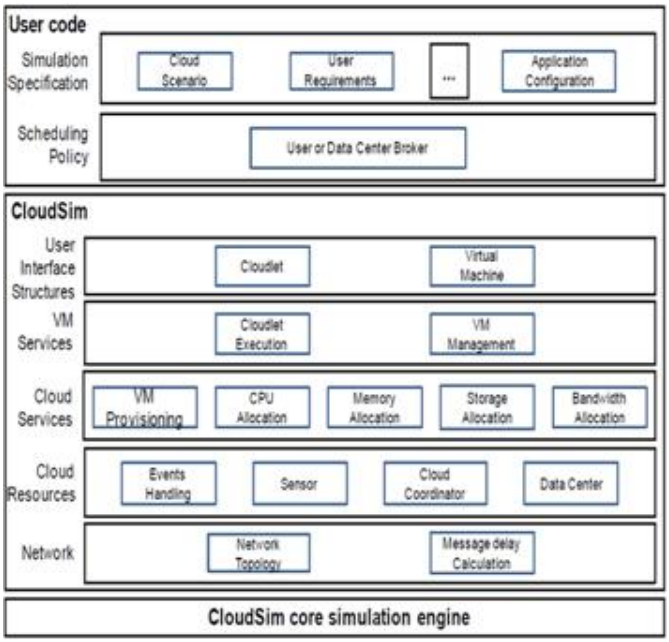

Fig. 5. Architecture of Cloudsim

This layer of CloudSim has responsibility to make communication possible between different layers.

b) Cloud Resources:
This layer includes different main resources like data-centers, cloud coordinator in the cloud environment.

c) Cloud Services:
This layer includes different service provided to the user of cloud services. The various services of clouds include IaaS,PaaS, SaaS.

d) User Interface:
This layer provides the interaction between user and the simulator. As the CloudSim has major limitation of lack of Graphical User Interface, so several variants has been developed such as CloudAnalyst, CloudReport, Green Cloud, etc

2) CloudAnalyst:
It was derived from cloudsim and extends some of its capabilities and features proposed. CloudAnalyst can be applied to examining behavior of large scaled Internet application in a cloud environment . The main feature provided by this tool is load balancing.

3) CloudReport:
This is a graphic tool that simulates distributed computing environments based on the Cloud Computing paradigm. It generates the simulation result in the form of HTML files. GreenCloud is a sophisticated packet-level simulator for energy-aware cloud computing data centers with a focus on cloud communications. It focuses on communication within a cloud that is all communication processes are simulated on packet level.

4) GDCSim:
GDC is a Green Data Center Simulator. It combines both modular and large scale entities.

5) Cloud Analyst:
Cloud Analyst is the most popular visualized type of simulators. This tool can be used easily and produces output in graphical format. It creates difference between programming environment and simulation environment.

6) Network Cloud:
It provides an extension to CloudSim by implementing network layer. It increases the performance of CloudSim.

7) MDCSim:
It is used to analyze and predict the hardware related issues of the servers and data centers.

8) SPECI Sim:
SPECI is Simulation Program for Elastic Cloud Infrastructure. It is used to analyze the scalability and performance concepts related to data centers.

9) DC Sim:
DC is Data Center Simulator, offering IaaS service of cloud and used to develop datacenter techniques.

10) GroudSim:
provides the IaaS service of the cloud in simulation tool and may also be used to provide PaaS and SaaS services of the clouds.

11) UEC:
Ubantu Enterprise Cloud (UEC) is an open stack public cloud. UEC is the most popular cloud service provider of the Unix. It works with the integration of different software.

12) iCanCloud:
This cloud does not require any modifications when there is requirement to test cloud in different architectures.
Network CloudSim extends the CloudSim tool by including network communication features in it.

B. Advantages of Cloud Simulation Tools

1) No Capital Investment Involved:
Simulation tools does not requires any installation and nor even maintenance cost.

2) Provides Better Results:
Simulation tools helps user to change input very easily as when needed, which provide better results as an output.

3) Risk are evaluated at earlier stage:
Simulation tools involve no capital cost while running as in case of being on cloud. This helps in identifying of risks with design or any parameter at earlier stage.

V. Security of Cloud Computing

The security of the user data is crucial when adopting a cloud computing model. Several methods have been implemented in securing user data such introducing encryption algorithms. Service traffic hijacking creates a major concern on using cloud computing for this allows the hacker to access the credentials of the genuine user hence he or she can eavesdrop on the users activities and transactions, manipulate user data, return false information and redirect your client to illegitimate

Fig. 6. Service Traffic Hijacking

sites. An important issue for cloud computing is the observation of security, which is beyond the basic technical details of security solutions.There are so many security challenges and threats in cloud computing. These are some of them. Abuse and nefarious use of cloud computing, malicious insiders, shared technology weaknesses, data harm and service traffic hijacking and etc.

Attribute Based Encryption is a technique which we can use to secure consumer data. In encryption technique there are some advanced encryption algorithms that can be applied to the cloud computing to increase the protection of privacy.Another encryption method to secure data in a cloud environment RSA encryption method. RSA encryption method uses a cryptographic algorithm where the encryption key is public and varies from the decryption key which is kept top-secret. DES is another method where it uses symmetric key for both encryption and decryption.RSA is asymmetric encryption and decryption algorithm. How this algorithm does the magic is it encrypts the user data for security purposes so that only the genuine authorized user can access the data. Data will be stored on the cloud and depending on the user requests the data will be delivered. In this scenario the public key is known to everyone but the private key is known only by the authorized user. Data encryption standard known as DES encrypt data within blocks with the size of 64 bits each. This produces 64 bits of cipher text. And the same key is used for encryption and decryption which the size of the key of this algorithm goes up to 56 bits.In cryptography a fully homomorphism encryption scheme is used. It allows data to be processed without being decrypted.

A. Service Traffic Hijacking Process

Account hijacking is carried out by the stolen credentials of the genuine user. Using the credentials the hacker can access sensitive data and manipulate data as per his likeness. Service traffic hijacking involves in hacker eavesdropping on activities, manipulating data, accessing data and returning falsified information.There are three states where the security breach can be occurred.

1) Transmission of sensitive data to the cloud server.
2) Transmission of sensitive data from cloud server to the clients computer.
3) The storage of sensitive data of the clients on the cloud.

In Fig 5, the left most side picture is where the genuine user enters the credentials to log in to the cloud server. This is where the intruder hacks and retrieve or eavesdrop on the activities and uses the sensitive data.

B. Prevention of Service Traffic Hijacking

There are few alternatives to be used to prevent service traffic hijacking. Observing user behavior can help to identify suspicious activities. Proactively monitoring user behavior detect unusual events such as downloading massive amount of data in a short period of time. Blocking the account for a period of time when suspicious activity occurs helps the genuine user to save his sensitive data. Hijacker needs two authentications to enter in to the user information. One authentication will not satisfy the requirements to enter thus this way hacker would not be able to penetrate the system and manipulate sensitive data. Prohibiting the sharing of the credentials between user and the service closes the door to hijackers on stealing the account credentials. This is where hijacker can easily access and retrieve the credentials. Understanding cloud provider service polices as well as service level agreements can help to reduce the threats.

C. Encryption In Cloud Computing

Nowadays, cloud computing acts a very important role in modern IT technology. But there are so many security challenges and threats in cloud computing. As a solution for this cloud data encryption mechanism is introduced. There are certain steps of process of data encryption .The data

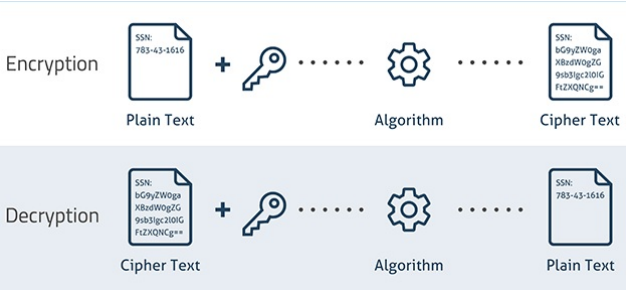

Fig. 7. Encryption-decryption process

pass through a mathematical formula called algorithm which converts it into encrypted data called cipher text. Encapsulate the message with key create a key by these algorithms .There are two types of encryption and they are asymmetric and symmetric.Firstly we talk about asymmetric encryption. There are two mathematically-related keys which are used. In public key (asymmetric) encryption: one to encrypt the message and the other to decrypt it.

Fig 7, describes the process of the encryption where the sub processes creates the cipher text and how the key is been used. It is required that the sender and receiver have a way to exchange secret keys in a secure manner when using this form of encryption.There are so many advantages in encryption. As one of a major advantage can take that encryption protect the cloud data completely. After encrypting the data it is very difficult to decode the information. And also provides the security for the encrypted data during transmission.

VI. Conclusion

Cloud computing is a new technology wide studied in recent years. Currently there are several cloud platforms that are employed in each in trade and in educational. The way to use these platforms could be a huge issue. During this paper, we have a tendency to delineate the definition, styles, and characteristics of cloud computing, cloud computing services ,storage,simulation tools and security of cloud computing. There are several issues in cloud computing. As an example of cloud computing issues is ability, Performance, Service Level Agreement (SLA), knowledge Confidentiality and measurability, knowledge Integrity, Load equalization, Synchronization in numerous clusters in cloud platform, and standardization, the protection of cloud platform

References

[1] A. Soofi, M. Khan and F. Amin, “Encryption Techniques for Cloud Data Confidentiality”, International Journal of Grid and Distributed Computing, vol. 7, no. 4, pp. 11-20, 2014

[2] Rachna Arora, AnshuParashar, “Secure User Data in Cloud Computing Using Encryption Algorithms”, International Journal of Engineering Research and Applications, vol. 3, no. 4, pp. 1922-1926, 2013

[3] YashpalsinhJadeja; Kirit Modi, “Cloud Computing - Concepts, Architecture and Challenges in Proceeding of International Conference on Computing, Electronics and Electrical Technologies [ICCEET], 2012.

[4] Qi Zhang, Lu Cheng and Raouf Boutaba, “Cloud computing: state-of-the-art and research challenges, Journal of Internet Services and Applications, May 2010, Volume 1, Issue 1, pp 718.

[5] Tharam Dillon, Chen Wu and Elizabeth Chang, “Cloud Computing: Issues and Challenges, in Proceeding of 2010 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 27-33, 20-23 April 2010.

[6] Yashpalsinh Jadeja, Kirit Modi, “Cloud Computing - Concepts, Architecture and Challenges in Proceeding of International Conference on Computing, Electronics and Electrical Technologies [ICCEET], 2012.

[7] Xu Wang, Beizhan Wang, Jing Huang, “Cloud computing and its key techniques in Proceeding of IEEE International Conference on Computer Science and Automation Engineering, 2011.

[8] Shyam Patidar Dheeraj Rane, Pritesh Jain, “A Survey Paper on Cloud Computing in proceeding of Second International Conference on Advanced Computing & Communication Technologies, 2012.

[9] Parveen Kumar, Anjandeep Kaur Rai, “An overview and survey of various Cloud Simulation Tools”, Journal of Global Research in Computer Science, January, 2014.

Survey on Web Usage Mining

Fasil P S, Reshma M and Silpa Raghavan
Vidya Academy of Science of Technology
Thrissur - 680501, India

Aparna S Balan
Vidya Academy of Science of Technology
Associate Professor of Computer Application
Thrissur - 680501, India
(email: aparna@vidyaacademy.ac.in)

***Abstract*—The process of finding out valuable knowledge drawn out from web data is known as web mining. Identifying the various patterns and utilizing the vast knowledge extracted from those patterns is important from various perspectives such as business intelligence, e-learning, personalization etc. The web mining area which deals with extraction of patterns from users weblogs is called as web usage mining, which is an implementation part of data mining. This paper focuses on the working of web usage mining, data sources for web usage mining and applications of web usage mining is explained in detail in this paper. Further, we explain the issues and current challenges in web usage mining**

***Index Terms*—OS, system, domain**

I. Introduction

With the advent of various technologies, social networking sites, e-business site, e-learning sites etc. Every second new data is generated. The data growth rate is exponential. On a single tweet, for example, one can find millions of messages or retweets on Twitter. All this is possible because of the internet. Thus, The World Wide Web is a large source of varied data, a collection of billion documents, which either comes from the content in the web pages or from the various hyperlinks or structure of the websites i.e. web structure or from the log files depicting the web usage. The data or information is collected from various points or data sources. The Internet is also a form of network and in a network data is available at various nodes. Here, nodes can represent servers, client machines, intermediate devices popularly called as proxy servers or various databases that are stored on machines. Web mining is the part of data mining from which we derive meaningful and useful knowledge from text or contents present in web pages, hyperlinks and user usage logs For a clearer understanding, web mining has 3 parts: Web Structure Mining, Web Content Mining, and Web Usage Mining. The Web Content mining deals with the raw data available in the web pages; the source data mainly consists of the textual information, images, graphic, audio etc. present in the web documents. Since web content mining is basically related to the web content, the main application fields of web content are a content-based ranking of web pages and contentbased categorization.

II. Web Usage Mining

Data mining is implemented in many forms, web usage mining is one of them. Whenever a user interacts with the web pages, Usage mining does the task of finding the hidden important information about user behavior, its page surfing patterns and other valuable information which is used for various purposes. Data is obtained from various sources for web usage mining. Some of them are discussed below:

A. Server Logs

These include the log files on the server-side for collecting information about the user such as IP address, access time, links visited etc[3]. Out of many weblogs, some acknowledged logs are as follows:

1) Common log format(CMF)
2) Explicit user input (EUI)
3) Cookies
4) Click streams

B. Client Data

This is the second data source which is collected from the host that accesses the websites or requests for the web pages. Remote agents are implemented in Java or javascript to access the data. They collect information like users navigational history. It is more reliable than server data as it has no problem with web caching and IP misinterpretation.

C. Intermediate Data

When the users request is granted through proxy servers or intermediate devices, intermediate data is generated. Therefore, proxy servers and packet sniffers also act as a source of data.

1) Proxy server
2) Packet sniffers

III. Web Usage Mining Process

The process of web usage mining comprises of the following steps:

- Data cleaning
- Data pre-processing
- Pattern Analysis
- Pattern discovery

A. Data Collection

The first step is a collection of data from varied sources. Apart from minor sources, the major sources are as follows:

1) The Server Side: The server-side has a substantial amount of information, which it represents in a standard format like CLF or ELF, stored in server log files or the database log files.

Issues: The major issue is user session identification i.e. identifying the click stream and the path followed by the user during visiting that website [3]. Even if we use cookies to remove this issue, the tracing of back button navigation is not performed at the server level, which makes it difficult to find the exact path followed by the user during its web browsing session

2) The Proxy Side: These are intermediate servers which improve navigation speed through caching and they collect data from huge servers by making groups of users. The collections of log data from proxy servers are similar to server-side data collection in many respects.

Issues: since there may be frequent caching between users and proxy servers it is difficult to reconstruct the session as the path of all the users navigation cannot be identified

3) The Client Side: The client-side data is the log files present on the user side. Here, we can track the usage data on the client side. JavaScript, Java applets and modified browsers are used to trace the data on the client machine. The problems of user identification and session identification are dodged.

Issues: Every machine may not fully support Javascript or Java applets to track the data.

B. Data Preprocessing

After collection of huge amount of data, it is necessary the data is consistent, integrated and relevant. Thus, Data needs to be ready for pattern discovery and analysis which is known as pre-processing of data. It is a time consuming and intricate process. It includes four different tasks: Data Cleaning, session identification and user identification with the rebuilding of users session, Recovering the information pertaining to the content of the page and structure and the data formatting.

1) Data Cleaning
2) User and Session Identification

1) Data Cleaning: The data collected on the internet has information which is not used as a request for graphical page content (images in .jpg and .gif format), ads etc [3]. Elimination of useless data from weblogs is the primary job of this step. A weblog is a website that has series of entries (logs) of the pages or sites or links surfed by the user. This is performed by spiders and robots [3].It is easier to drop graphical requests but the navigation patterns of robots and spiders are explicitly identified and removed. By keeping a track of access to robots.txt file these patterns can be removed effectively [3,4]. Another approach used for robots that use false user agent in HTTP session from actual approach is heuristic based The classifier is trained with the help of well-known robots navigational path and the further categorization is done using the acquired

C. User and Session Identification

In this step, we find the users session from the weblog data obtained and as the next step within the identified session, rebuilding of users navigation route is performed. For session identification, we use cookies, URL rewriting etc.

1) User Identification: It is the process of identifying the users on the web i.e. which user accesses the web pages or requests for the websites or web pages. There are various approaches to automating user identification. A unique user id can be given to each user in a log file; this is one of the methods of identifying the user. However, if the user uses same machines or proxy servers it is difficult to identify the user. Cookies are used for finding the user but are subject to deletion of disability. Another method is of the various special services present on the internet like in or fingered services for user recognition Even if the IP of the user is same, by analyzing the various weblogs the difference in browser types, operating systems, and their versions are identified.

The topology of the site can be merged with access log entries to identify if the user is new or not. This is distinguished by the fact that if access with an identical IP is made to an identical page without a direct hyperlink between them, then the user is a new user. Instead of cookies, the web server includes a unique ID in the URL and the user creates a bookmark for one of the delivered pages in order to gain access to the website. Thus, this method is not fully automatic as user needs to bookmark the

2) Session Identification: The information of the navigation pattern of the user is put into code by session identification. User sessions can be found out with the help of various proposed method which is either based on time or based on the content.

IV. TECHNIQUES

There are various techniques for the analysis of web usage data and discovering knowledge from it. The major techniques are as follows:

1) Statistical Analysis
2) Assosiation Rules
3) Clustering
4) Classification
5) Sequential Patterns
6) Dependency Modeling

V. APPLICATIONS

Valuable patterns are drawn out from the outcomes produced from the web usage mining process. The applications of results are summarised below:

A. Personalization of Web Content

Personalizing the experience of the user is important from the business perspective of many web-based applications. Personalization means showing those results to user, which the user wants to see based on his navigation patterns, search patterns and other patterns discovered from web log data[1]. For example, through users browsing pattern we can identify

the type of web pages and websites the user visits and customize the web for him using this analysis.

B. Pre-fetching and Caching

The server response time, access time is cut using the caching and pre-fetching technique by means of using the outcomes extracted by analysis of weblogs.

C. Support for the Design

To make the websites efficient and flexible to use by all the users, weblog extracted data patterns helps in design issue of the wesite.

D. To Improve Customer Satisfaction

With personalized experience the user gets a better access to the website and an improved experience on its features, hence improving customer satisfaction.

E. Business Intelligence

Drawing out of knowledge from information is a primary task performed by web usage mining and determines customer behavior. It helps in determining effective marketing strategies that help in increasing the sales [10,28].Business intelligence is all about helping people make good decisions and maintaining competitiveness in the marketplace.

F. Enhanced E-learning

The courses activities can be traced using the web usage data and suggestions for system improvement are placed using this information.

VI. Current Issues and Challenges

The major issues in web usage mining are discussed below: Privacy is considered as one of the most major issues in web usage mining. Since data is collected from vast and varied data sources like cookies, weblogs, URLs etc. it is difficult to support the privacy of data. The solution to privacy problem is P3P i.e. privacy preference standard which proposes various privacy standard format [12,26]. Another problem is dealing with large and vast volumes of data which is exponentially growing with time. The lack of information on the comparison of various tools which makes evaluation criteria difficult

VII. Future Trends

Web usage mining has various issues which give open research areas in this field which help in developing future trends in this domain. There are two prominent issues in this area as discussed above. Firstly, privacy is a big challenge. Secondly, integration of semantics within websites, this is also an open research area, which is also a research area [3]. There are several approaches followed to develop trends like modeling. The future trends will mostly revolve around privacy and semantics. Thus, we have various open research area in web usage mining which gives rise to future trends.

VIII. Conclusion

Due to vast extensions of the network, various websites and applications have emerged that keep the extensive amount of user information. Web usage mining helps in discovering the patterns from users weblogs and through that extracted information, knowledge is drawn out and used in various fields like e-commerce, online business sites etc. Studying and analyzing this information helps the designers in catering to users specific needs i.e. personalizing user experience and efficiently organizing the websites. Also, customer behavior patterns help in establishing business rules and based on these the organization identifies the future needs the customer may develop and their current likings.In this paper, we presented a survey on web usage mining, the processes involved, the techniques used. Further, we have discussed various issues and challenges and the future trends which are mainly related to privacy and dealing with large volumes of data

References

[1] Guerbas A, Addam O, Zaarour O, et al. “Effective web log mining and online navigational pattern prediction”, Journal of Knowledge-Based Systems, 2013, 49:50-62.

[2] Wang Sibao, Li Yinsheng, “Mining User Preferred Browsing Paths Based on Web Logs”, Journal of Computer Applications and Software, 2012, 29(1):164-167.

[3] Fu Zhitao, “Research and implementation of network user clustering based on web log”, Nanjing University of Science and Technology, 2007. DOI:10.7666/d.y1153993.

[4] Aghabozorgi S R, Wah T Y, “Using Incremental Fuzzy Clustering to Web Usage Mining”, Journal International Conference of Soft Computing and Pattern Recognition, 2009:653 - 658.

[5] Chen Jian, Yin Jian, “Research on Key Technology in Web Usage Mining”, Journal of Computer Engineering, 2005, 31(9):4-6. DOI:10.3969/j.issn.1000-3428.2005.09.002.

[6] Sudheer Reddy K, Kantha Reddy M, Sitaramulu V, “An effective data preprocessing method for Web Usage Mining”, International Conference on Information Communication and Embedded Systems, 2013:7 - 10.

[7] Pang-Ning Tan and Vipin Kumar, “Modeling of web robot navigational patterns”, In WEBKDD 2000 - Web Mining for E-Commerce Challenges and Opportu- nities, Second International Workshop, August 2000. 142.

A Study on Web Content Mining

Fathima Mol, Syamdev A J and Tony Tom
Vidya Academy of Science & Technology
Thrissur - 680501, India

Aparna S Balan
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: aparna@vidyaacademy.ac.in)

***Abstract*—Web Mining is extracting information from the web re-sources and finding interesting patterns that can be useful from ever expanding database of World Wide Web. Whenever we talk about data, we conclude that there is a huge range of data on World Wide Web. Due to heterogeneity and unstructured nature of the data available on the WWW, Web mining uses various data mining techniques to discover useful knowledge from Web hyperlinks, page content and usage log. Web Content Mining is a component of Data Mining. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the Web to the user requesting the information. This paper deals with a preliminary discussion of Web content mining, contributions in the field of web mining, the prominent successful tools and algorithms.**

***Index Terms*—Web content mining, structured data mining, unstructured data mining, semi-structured data mining.**

I. Introduction

WEB is the largest data source in the world. Web mining aims to extract and mine useful knowledge from the Web. It is a multidisciplinary field involving data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc. The amount of information on the Web is huge, and easily accessible. The knowledge does not come only from the contents of the web pages but also from the unique feature of Web, its hyperlink structure and the diversity of contents. Analysis of these characteristics often reveals interesting patterns and new knowledge which can be helpful in increasing the efficiency of the users.Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques. Webpage can be in traditional text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents.

II. Web Content Mining

Web Content Mining is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. It includes extraction of structured data/information from web pages, identification, similarity and integration of datas with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [3].

III. Web Content Mining Strategies

Web Content Mining Approaches: Two approaches used in web content mining are Agent based approach and database approach [4],[5]. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, and personalized web agents [6]. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Web content mining has the following approaches to mine data :

- Unstructured text mining
- Structured mining
- Semi-structured text mining
- Multimedia mining.[8]

A. Unstructured Text Data Mining

Most of the Web content data is of unstructured text data. Content mining requires application of data mining and text mining techniques [7]. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are listed below:

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Information visualization

B. Structured Data Mining

The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are given below:

- Web crawler
- Wrapper generation
- Page content mining

C. Semi-Structured Data Mining

Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intradocument structure. The techniques used for semi structured data mining are given below:

- Object Exchange Model (OEM)
- Top Down Extraction
- Web Data Extraction language

D. Multimedia Data Mining

The techniques of Multimedia data mining are:

- SKICAT
- Color Histogram Matching
- Multimedia Miner and Shot Boundary Detection

IV. Web Content Mining Algorithms

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to fetch the information are described

1) Decision tree:
   The decision tree is one of the powerful classification techniques. Decision trees take the input as its features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5.
   The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [12].
2) k-Nearest Neighbour:
   KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation [15].
3) Naive Bayes:
   Naive Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP).Given a problem with K classes $\{C1, \ldots, CK\}$ with so called prior probabilities $P(C1), \ldots, P(CK)$, can assign the class label c to an unknown example with features such features $x = (x1, \ldots, xN)$ such that $c = \arg\max P(C = ck|x1, \ldots, xN)$, is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, that is choosing the class with the maximum a posterior probability given the observed data. This posterior probability observed data. This posterior probability can be formulated:

   $$P(C = ck|x1, \ldots, xN)$$
   $$= p(C = c)P(x1, \ldots, xN|kC = c)P(x1, \ldots, xN)$$

   As the denominator is the same for all classes, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the accessible classes. This may be quite difficult taking into account the dependencies between features. This approach is to assume conditional independence i.e. $x1, \ldots, xN$ are independent. This simplifies numerator as $P(C = c)P(x1k|C = c).P(xNk|C = c)$, and then choosing the class c that maximizes this value over all the classes $c = 1, \ldots, K$. [12].
4) Support Vector Machine:
   Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear data and uses a nonlinear mapping to transform the original training data into a higher dimension. Among the new dimension, it searches for the linear optimal separating hyper plane (i.e., decision boundary). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyper plane [15].
5) Neural Network:
   The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer. Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [12]. As network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.
6) Cluster Hierarchy Construction Algorithm (CHCA):
   The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence

of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed). Using the reduced table, we create a cluster hierarchy by examining each row, starting with those with the fewest terms (fewest number of 1s); these will become the most general clusters in our hierarchy.

The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster. Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above.

If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it. This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created.

The next step in the algorithm is to assign the web pages to clusters in the hierarchy. In genera there will be some similarity comparison between the terms of each web page (rows in the original table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page.

Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

V. Web Content mining Tools

Web content mining tools helps to download the essential information. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda and Web Content Extractor, Rapid Miner.

A. Rapid Miner

Rapid Miner is open source software and it is a tool for extracting information from web, Contains inbuilt algorithm. It can generate algorithm by itself.

Features:

- Easy to use.
- Reduce time.
- Open source software.

B. Screen-scaper

Screen-scraping is a tool for extracting/ mining information from web sites [11]. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. Features: Screen-scraper present a graphical interface allowing the user to designate URLs, data elements to be extracted and scripting logic to traverse pages and work with mined data.

Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked. This also facilitates scraping of information at periodic intervals. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time after which the results are displayed in a single interface.

C. Automation Anywhere

It is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining [14].

Features:

- Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly. Web record and Web data extraction.

D. Web Info Extractor

This is a tool for data mining, extracting Web content, and Web content analysis. It can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

Features:

- No need to learn boring and complex template rules and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file
- Can deal with Web page in all language
- Running multi-task at the same time
- Support recursive task definition.

E. Web Content Extractor

It is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet[13].It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner. This tool allows users to extract data from various websites such as online stores, online actions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source.

Features:

- Helps to extract/collect the market figures, product pricing data, or real estate data.

- Helps users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- Assists to Journalists extract news and articles from news sites.
- Helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

F. Mozenda

This tool enables users to extract and manage Web data [15]. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, the data to be used in other applications or as intelligence.

Features:

- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence.

VI. Conclusion

The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in web content mining but this paper tried to expose the various web content mining strategy and explore some of the techniques. Then we described some tools web content mining.

References

[1] errouz, A., Khentout, C., Djoudi, M., "Overview of Visualization Tools for Web Browser History Data", IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).

[2] Raymond Kosala, Hendrik Blockeel, "'Web Mining Research: A Survey", ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[3] Han, J., Kamber, M. Kamber, "Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

[4] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", in Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668677,1998.

[5] R. Cooley,. B. Mobasher, J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web, in Proceedings of Ninth IEEE International Conference. pp. 558 567, 3-8 Nov. 1997.

[6] Inamdar, S. A. and shinde, G. N., "An Agent Based Intelligent Search Engine System for Web Mining", International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[7] V. Bharanipriya & V. Kamakshi Prasad, "Web Content Mining tools: A Comparative Study", in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[8] Johnson, F., Gupta, S.K., "Web Content Minings Techniques: A Survey", International Journal of Computer Application. Volume 47 No.11, p44, June (2012).

[9] Dunham, M. H., "Data Mining Introductory and Advanced Topics. Pearson Education, 2003.

[10] R. Baeza-Yates and e. Berthier Ribeiro-Neto, Modern"'' Information Retrieval", Addison-Wesley Longman Publishing Company, 1999.

[11] screen-scraper, [Online] Avilable: `http://www.screen-scraper.com`.

[12] Darshna Navadiya, Roshni Patel, "Web Content Mining Techniques-A Comprehensive Survey", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue10,December- 2012 ISSN: 2278-0181

[13] Web Content Extractor help. [Online] Available: `http://www.newprosoft.com/web-content-extractor.htm`

[14] Automation Anywhere Manual. AA, [Onlinne] Available: `http://www.automationanywhere.com` Viewed 06 February

[15] Zenda [Online] Available: `http://www.mozenda.com/web-mining-software`

[16] Zhang, Q., Segall, R.S., "Web Mining: A Survey of Current Research, Techniques, and Software", International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

Intrusion Detection and Prevention System in Cloud Computing

Fila Jose, Gopika K and Silpa P R
Vidya Academy of Science & Technology
Thrissur - 680501, India

Siji K B
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: siji.k.b@vidyaacademy.ac.in)

***Abstract*—Cloud Computing has emerged as a model to process large volumetric data. Though Cloud Computing is very popular, cloud security could delay its adoption. Security of the cloud must provide data confidentiality and protection of resources. The security of Cloud Computing extends the physical security which securing equipment constituting the Cloud architecture, data security must ensure that the service to the client, and network security which plays an important role in ensuring service and reliable protection against attacks and intrusions. At this level, security systems operating in traditional networks are also used in the new model to strengthen its network security. Here we explain about the Intrusion Detection(IDS) and Prevention (IPS) System in cloud security(IDPS).**

***Index Terms*—Cloud computing, intrusion detection, intrusion prevention, smart grid**

I. Introdution

Over the last decade, our society has become technology dependent. People rely on computer networks to receive news, stock prices, email and online shopping. The integrity and availability of all these systems need to be defended against a number of threats. Amateur hackers, rival corporations, terrorists and even foreign governments have the motive and capability to carry out sophisticated attacks against computer systems (Choo, 2011). Therefore, the field of information security has become vitally important to the safety and economic well-being of society as a whole. The rapid growth and widespread use of electronic data processing and electronic business conducted through the massive use of the wired and wireless communication networks, Internet, Web application, cloud computing along with numerous occurrences of international terrorism, raises the need for providing secure and safe information security systems through the use of firewalls, intrusion detection and prevention systems, encryption, authentication and other hardware and software solutions.

Cloud Computing is one of todays most promising technologies due to its cost-efficiency, flexibility and scalability for computing processes. However, the complex architecture of cloud infrastructure and the different levels of users lead to special requirements especially in security area. The Cloud provider is responsible for providing secure, reliable and trustful services to its consumers. Network intrusion detection system and network intrusion prevention system (IDPS), is a pioneer active security-defensive mechanism that is ideal to be used in cloud computing.

II. Attack Types

We consider two types of attacks: portscan and distributed portscan.

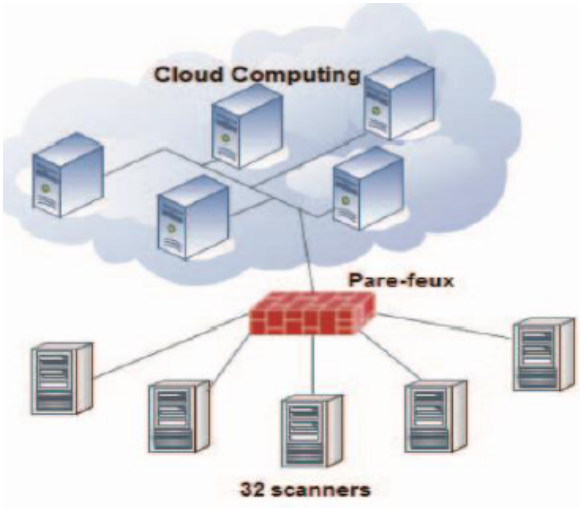

Fig. 1. Types of attack in cloud computing model

A. Portscan

Fyodor, initial contributor of the audit tool Nmap describes the port scan as a recognition phase. During which an attacker determines what types of network protocols or services offers a machine to discover exploitable communication channels in order to establish his attack. In general, the port scan is an exchange of network packets. According to it, the attacker can enter the port status: open, closed or filtered.

Variables of a Portscan: When an attacker wants to perform a portscan on a network, he can adjust some variables. These variables let the attacker be more discrete or know specific information about remote scanned hosts.

1) Techniques:
   Depending on the goals of the attacker, he would use a specific technique rather than another. Most of portscan techniques give information about state of the targeted ports whereas other techniques give information about the service or the operating system.
2) Timing:
   An attacker can adjust the speed rate of portscan attacks. By doing this, he can evade IDSs, because most of them do not detect portscans when they are executed very slowly.
3) Targeted ports:
   Lee describes different types of targeted port. Her paper introduces vertical, horizontal and block scans, presented below. Vertical scans are portscan that target numerous ports on a singular remote host. That type of portscan tries to discover a weakness on a particular host. Contrary to vertical scans, horizontal scans target only one port on several remote hosts. This lets an attacker search a specific weakness on a given network. Horizontal scans are commonly used by worms, which are aware of a particular vulnerability. A block scan is a combination of vertical and horizontal scans; it consists in a portscan of several ports on several remote hosts.

B. Distributed Portscan

Distribution of an attack is to divide it into tasks that will be performed by different machines in many ways for these techniques are. For its methods of distribution, the attacker has a set of machines, called scanners used for his work. The authors in present two methods of distribution:

1) Nave distribution:
   It consists in a sequential distributed scan, the attacker selects one of the scanners he controls and starts scanning the target network with it. When the scanner is detected, the attacker selects a different scanner and resumes the portscan. The process continues until the portscan is completed. We expect this type of distributed portscan to be linear: if a scanner can scan x ports, two scanners should be capable to scan the double.
2) Parallel distribution:
   To change the distribution consists in splitting the whole set containing targets and ports between scanners. Each scanner has a sub-task to perform, and then he communicates the results to a coordinator. We expect this technique to overrule IDSs detection due to the generated traffic.

III. Intrusion Detection and Prevention Systems Taxonomy

Attacks that come from external origins are called outsider attacks. Insider attacks, involve unauthorized internal users attempting to gain and misuse non-authorized access privileges. Intrusion detection is the process of monitoring computers or networks for unauthorized entry, activity or file modification. An IDS is a software that automates the intrusion detection process and detects possible intrusions. An IDPS is a software or hardware device that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents. IPSs are differentiated from IDSs by one characteristic; IPS can respond to a detected threat by attempting to prevent it from succeeding. The IPS changes the attacks content and/or changes the security environment.

A. Intrusion Detection and Prevention Systems Model

Our network-based intrusion detection and prevention system (IDPS) consists of Pre-processing part, Classification part, and Protection part as shown in Figure 1. The system starts detecting packet from the Ethernet, and send packet data to the pre-processing part for extracting important features to form a data record within a certain time interval. Then the pre-processing data is sent to the Classification part to identify types of attacks, or else the data is normal network activity. The result from this part is then sent to the Protection part.

1) Pre-processing part:
   In this Pre-processing part, we use packet sniffer in Java language to capture packet information between a source-destination IP pair including IP header, TCP header, UDP header and ICMP header from the Ethernet interface card. The information is extracted from the packet header within a certain period of time (e.g. 1-2 seconds) and store as a record. We also record port numbers and package received time in this step.
2) Classification part:
   In Classification part, we take each of the pre-processed data records to classify it as normal data or attack data. The pre-processed data will be classified by machine learning algorithms written in java. A list of well-known algorithms consisting of Ripple Rule, Random Forest, Decision Tree C4.5, and Bayesian Network can be used for classification. If we select more than one algorithm in a time, majority voting will be applied to the result of classifier. The result from classification which could be Normal, Probe, or DoS will be saved into a log file and also sent to the Protection part.
3) Protection part:
   In Protection part, the network data packets will be blocked by using IP table if network attacks are detected. The system will get results from previous classification part for making decision on which operation to be used. If the result of network types is Probe, the system will record senders IP address as attacker IP and block or drop all packets from the attacker IP. If the result is DoS, the system will record that the connection port number was attacked and then block or drop all packets going through the attacked port number. The system does nothing if the result of classification is normal.

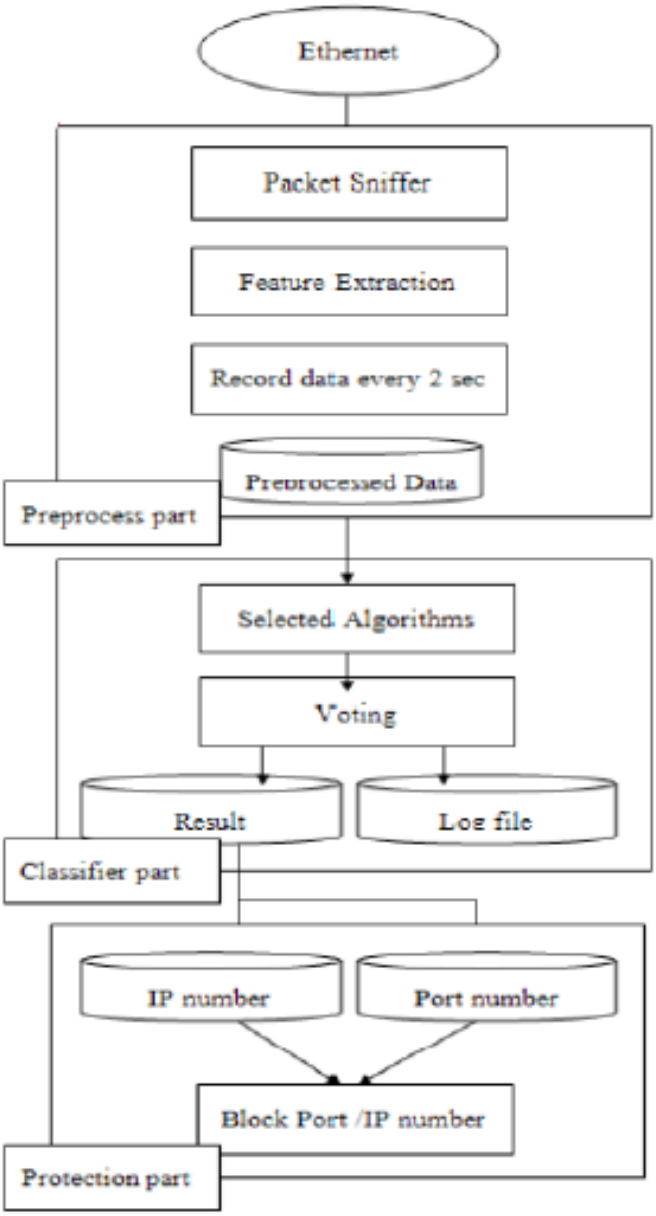

Fig. 2. Types of attack in cloud computing model

B. Functions

IDPSs serve four essential security functions: they monitor, detect, analyse, and respond to unauthorized activities. An IDPS detects intrusion by analysing the collected data. The monitored environment can be network-based, host-based or application-based.

1) Network-based (NIDPS):
Monitors network traffic for particular network segments or devices and analyses the network and application protocol activity to identify suspicious activity.
2) Host-based (HIDPS):
Monitors all or parts of the dynamic behaviour and the state of a computer system. Much as an NIDPS will dynamically inspect network packets, an HIDPS might detect which program accesses what resources.
3) Application-based (AIDPS):
Concentrates on events which occur in some specific application through analysing the application log files or measuring their performance. Its input is data sources of running applications.
The real-time IDPS can also be run for off-line analysis through historical data to identify past intrusions.

C. Intrusion Detection System

1) Misuse detection:
This method uses specifically known patterns of unauthorized behaviour, called signatures, to predict and detect subsequent similar attempts.
2) Anomaly detection
Designed to uncover abnormal patterns of behaviour. IDPS establishes a baseline of normal usage patterns, and whatever deviates from this get flagged as possible intrusions (Thatte et al., 2011). What is considered to be an anomaly can vary, but normally, any incident that occurs on frequency greater than or less than two standard deviations from the statistical norm raises an eyebrow (Bringas and Penya, 2009). There are various categories of anomaly detection proposed, but the three most used ones are as follows.
3) Statistical:
In this approach the system observes the activity of subjects (such as CPU usage or number of TCP connections) in terms of statistical distribution and creates profiles which represent their behaviors. Therefore, they make two profiles: one is made during the training phase and the other is the current profile during the detection. An anomaly is recognized if there is a difference between these two profiles.
4) Machine learning based:
This technique has the ability of learning and improving its performance over time. It tends to focus on constructing a system which can optimize its performance in a loop cycle and can change its execution strategy according to feedback information. System call-based sequence analysis, Bayesian network and Markov model are the most frequently used techniques.
5) Data mining based:
Data mining techniques can help to improve the process of intrusion detection by unfolding patterns, associations, anomalies, changes, and important events and structures in data. Classification, clustering and outlier detection, and association rule discovery are data mining techniques used in IDPS.
6) Hybrid:
This approach has been proposed to enhance the capabilities of a current IDPS by combining the two methods of misuse and anomaly. The main idea is that misuse detects known attacks while anomaly detects unknown attacks.

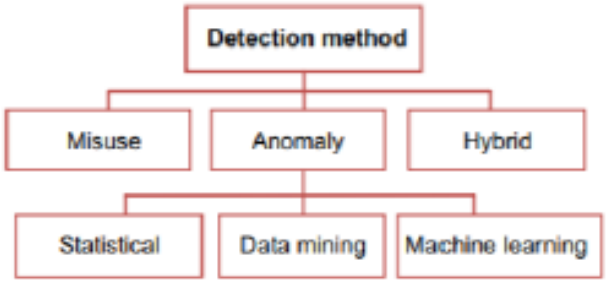

Fig. 3. Detection methods

D. Intrusion Prevention System (IPS)

IPS has been developed from IDS and contains the functionality of the latter, but they are more sophisticated with the ability to take immediate action to prevent malicious behaviour, in [11], the paper presents a comparative study between the two systems as shown in Table 1.

TABLE I
COMPARISION BETWEEN IDS AND IPS

IDS	IPS
Installed on network segments and on host	Installed on Network Segments and on host
Sits on network passively.	Sits inline(not passive).
Central management control.	Central management control.
Cannot parse encrypted traffic.	Better at protection application.
Better at detecting hacking attacks.	Ideal for blocking defacement

Some unknown attacks are formed from known attacks and can be produced by changing the signatures of captured packets. The integration of the signature algorithm Apriori will allow us to capture network packets with known signatures in the entrance and therefore generating new attack signatures in the output. These new signatures are then injected in our database of the IDPS to detect variants of unknown attacks like distributed attacks with a quick and efficient manner.

IV. A CLOUD COMPUTING BASED COOPERATIVE INTRUSION AND DETECTION AND PREVENTION SYSTEM FRAMEWORK

Cloud providers deploy virtualization to a physical infrastructure base to provide resources to users in an economical way. There are three main types of service that can be provided in a cloud environment: Software as a service (SaaS), Platform as a service (PaaS), and Infrastructure as a service (laaS). Each of these services defines a certain level of resources to be delivered. Talking about cloud computing refers to both the applications level services in the lower level up to the hardware in the data centres that provide those services in the upper level. The special about such an environment is that costumers or end users pay only for what they need regardless of the underlying storage or infrastructure.

Within this attention-grabbing environment, security issues arise with its entire constrains: privacy, confidentiality, integrity, authentication and availability. Similar types of attacks to those in regular computing networks can target a cloud computing infrastructure but with new vulnerabilities and weaknesses to exploit. Intrusion detection system (IDS) is a pioneer active security-defensive mechanism that is ideal to be used in cloud computing. IDS is a real-time monitoring system that is essential to detect or prevent intrusions before they actually take place. Network intrusion detection system and network intrusion prevention system (IDS/IPS) are sometimes combined in one term IDPS that defines IDS that afford the required prevention capabilities. Nevertheless, a single IDPS deployed independently in cloud regions, without any cooperation and communication, can easily suffer from problems such as single point of failure and low detection capabilities.

A. cl-CIDPS System Features and Architecture

A structured architecture for cl-CIDPS components is presented in Fig. This architecture integrates all major blocks that functions in each cl-CIDPS agents. The packet sniffing block is responsible of analyzing each received packets. These packets will be analyzed by both signature and anomaly based system. Several storages exist to facilitate the detection process: Signature List, IP-block List and Authentic List. Signature list enclose list of known and detected signatures. The IP block list contain list of malicious IP addresses. List of authentic IDPS agents addresses and keys are stored in the authentic list. Upon the detection, The Prevention block will decide the proper action. Moreover, the malicious packets are sent to the logging and alert system blocks.

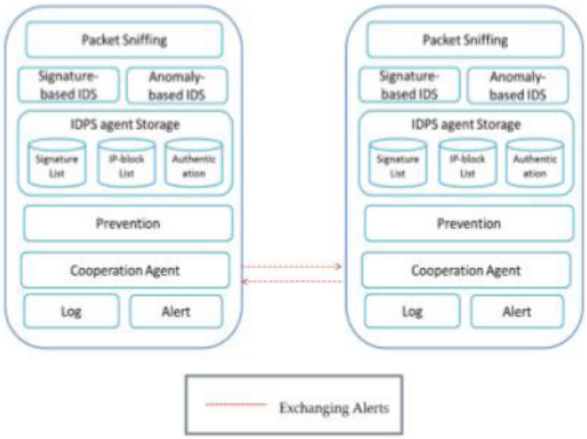

Fig. 4. cl-CIDPS main block and components

Cloud providers are responsible to ensure that users get the right security services and should insure that those services are not affecting the cloud benefits. For each deployed model (e.g. SaaS, PaaS, or IaaS), there exist trade-offs in integrated features, such as complexity vs. extensibility and security.

B. cl-CIDPS Main Blocks Methodology and Workflow

Within cl-CIDPS framework, IDPS agents are integrated in each edge node in every cloud region. When getting targeted by an intrusion message, the agent will send out the alert to other IDSs defined in its authentic list. All agents are required to exchanges alerts and have the ability to drop the message or prevent it from attacking other nodes in the network. After undertaking the correct action, the new signatures and IP addresses is added into the block table if the messages or the attacker are regarded as a new intrusions. Signature-based and anomaly-based detection modes are developed separately in cl-CIDPS framework. Each of the two detection modes has its own detection mechanism but similar communication procedure.

- cl-CIDPS Signature-Based Detection Mode:
  Signature-based detection mechanism is based on content inspection. It is considered more precise and accurate than anomaly based detection and has no false positive rate. Each individual signature-based cl-CIDPS agent is identical to the other and is communicating in a pure P2P fashion. Each individual signature-based cl-CIDPS agent is identical to the other and is communicating in a pure P2P fashion.
- cl-CIDPS Anomaly-Based Detection Mode:
  It is a general term that indicates detecting abnormal behaviours of the designated network. In cl-CIDPS framework, an anomaly-based IDPS agent can detects DoS attacks based on a pre-determined threshold. The threshold is computed based on the regular byte rate or bytes received per unit of time. This rate is computed for each received source IP-address. If the rate exceeded a normal value, the agent should drop the packet and log information.

V. Smart Grid

Smart Grid (SG) critical infrastructure systems are susceptible to high security risks cyber-attacks. It necessitates resilient and protective Intrusion Detection and Prevention Systems (IDPSs) to protect them. Since traditional signature and anomaly detection of intrusions are insufficient to make SGs safe, therefore a fully distributed managed Collaborative Smart IDPS (CSIDPS) is proposed. It is robust, flexible and scalable to satisfy the core requirements of IDPS for future SGs by including a set of autonomic, machine learning and ontology knowledge-base inference engine and fuzzy logic risk manager functionalities. In comparison to IDPS, CSIDPS increases detection accuracy and decreases false positive alarms. The main purpose of a cooperative intrusion detection system for clouding computing network to reduce the impact of DoS attack.

Smart Grid puts information and communication technology into electricity generation, delivery, and consumption. The security and encryption infrastructure elements are critical to enable Smart Grid.

A. Security in Smart Grid

Traditional security systems are not sufficient because of their lack of efficiency, their maintenance cost while deploying to the existing systems etc.

A Collaborative SIDPS (CSIDPS) is proposed in terms of a system structure and functionality. It provides robustness and seamless integrated protection within supply and demand sides of a Smart Grid to overcome large scale attacks and to use the computational resources efficiently.

The structure of a typical IDPS is based on two types: individual or collaborative. An individual IDPS is normally achieved by physically integrating it within a firewall. These IDPSs are ineffective in protecting critical infrastructure assets because they produce more irrelevant and false alarms.

A collaborative IDPS consists of multiple IDPSs over a large network where they intercommunicate, and are more efficient to detect and prevent intrusions. These IDPSs have two main functional components, the detection element and the correlation handler. Detection elements monitor their own sub-network or host individually and generate low level alerts. The correlation handler transforms these alerts into high level event reports. A collaborative IDPS has three structural forms:

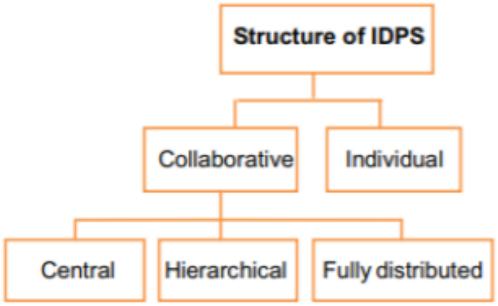

Fig. 5. Signature of IDPS

1) Centralized:
   Each IDPS acts as a detection element where it produces alerts locally. The generated alerts are sent to a central management control server that plays the role of a correlation handler to analyse them to make an accurate detection decision. The main drawback is that the central unit is extremely vulnerable, and any failure can lead to deactivating the whole correlation function.
2) Hierarchical:
   The IDPSs at the lowest level work as detection elements, while the IDPSs at higher levels are furnished with both a detection element and a correlation handler acting as aggregators. The IDPSs in the higher level correlate alerts from both their own level and lower levels. The correlated alerts are passed to a higher level for further analysis, an aggregation decision. This is more scalable than the centralized approach, but still suffers from the shortcomings of any function employed using

the centralized approach which can partly paralyze/stop the whole system operation.

3) Fully distributed:
There is no centralized coordinator to process the information, it compromises fully autonomous systems with distributed management control. All participating IDPSs have their own detection element and a correlation handler acting as an aggregator. Its advantages are that none of the IDPSs need to have complete information of the entire network topology; it has a more scalable design since there is no central entity responsible for doing all the correlation work; and simplifying the alarm correlation locally. The problem is that the information of all alerts is not available during the decision making which reduces detection accuracy.

VI. Conclusion

Cloud Computing is architecture full of resources. It offers several layers of services according to the needs of users, the security problems only delay its widespread adoption. Lately, attackers use several tools and solutions to exploit the vulnerabilities of information systems such as network faults or weaknesses of a machine or application specific cause then data loss and/or resources of the target environment. Cloud Computing and multi-path side, attacking take advantage of this feature to generate attacks and intrusions in a distributed manner and discreet. Until now, security solutions are not yet adapted to this new concept. Indeed, in such an environment, the more customers and paths, the greater the intrusion is effective.

Designing intrusion detection and prevention system for cloud environment require advance knowledge in network management, security aspects and communication protocols. This work attempts to encapsulates all these trades and propose a system that can be integrated in the cloud infrastructure level. cl-CIDPS was evaluated within a simulated cloud environment and proved its efficiency in detecting attacks and exchanging alerts. cl-CIDPS is built on the cloud infrastructure level deploying a pure peer-to-peer communication protocols.

For future work, unsupervised learning algorithms and new techniques will be considered together as a hybrid IDPS approach to enhance the performance of anomaly intrusion detection.

References

[1] M. Sabhnani and G. Serpen, “Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, Inter Conference: Machine Learning, Models, Technologies and Applications (MLMTA), 2003, pp. 209-215.

[2] K. Labib and R. Vemuri, “NSOM: A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps, Networks and Security, 2002.

[3] Weka library, “Data Mining Software in Java” [Online] Available: `http://www.cs.waikat o.ac.nz/ml/weka/`.

[4] Portscan techniques [Online]. Available: `http://nmap.org/book/man-portscanning-techniques.html`.

[5] Zargar, S.T., Takabi, H., Joshi, J.B., “DCDIDP: a distributed, collaborative, and data-driven intrusion detection and prevention framework for cloud computing environments”, in IEEE 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 332341 (2011).

[6] Qu, X., Liu, Z., Xie, X., “Research on distributed intrusion detection system based on protocol analysis”, in IEEE ASID 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, Hong Kong, pp. 421424 (2009).

[7] Lo, C.C., Huang, C.C., Ku, J., “A cooperative intrusion detection system framework for cloud computing networks”, in IEEE 39th International Conference on Parallel Processing Workshops (ICPPW), San Diego, pp. 280284, September 2010.

[8] Roschke, S., Cheng, F., Meinel, C., “Intrusion detection in the cloud”, in DASC2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Chengdu, pp. 729734 (2009).

[9] Bye, R., Camtepe, S.A., Albayrak, S., “Collaborative intrusion detection framework: characteristics, adversarial opportunities and countermeasures”, in Usenix Workshop on Collaborative Methods for Security and Privacy, CollSec, USENIX Association, August 2010.

[10] Luther, K., Bye, R., Alpcan, T., Muller, A., Albayrak, S., “A cooperative AIS framework for intrusion detection” in ICC2007 IEEE International Conference on Communications, pp. 14091416. IEEE, Glasgow (2007).

Computer Clusters

Haritha P M, Maya K and Sharafudheen K M
Vidya Academy of Science & Technology
Thrissur - 680501, India

Salkala K S
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: salkala@vidyaacademy.ac.in)

***Abstract*—A cluster of computers joins computational powers of the compute nodes to provide a more combined computational power. Therefore, as in the client-server model, rather than a simple client making requests of one or more servers, cluster computing utilize multiple machines to provide a more powerful computing environment perhaps through a single operating system. A HPC system is described by numerous processors, heaps of memory, fast systems administration, and expansive information stores all common crosswise over numerous rack-mounted servers. This paper deals with the input/output techniques, flow-control mechanism, storage architecture, slot allocation, task computing on clusters and massive data computing framework.**
***Index Terms*—I/O Architecture, I/O Configuration, static flow-control, solid-state-drive, Dynamic scheduling**

I. Introduction

A COMPUTER cluster is a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system. Unlike grid computers, computer clusters have each node set to perform the same task, controlled and scheduled by software. Grid computing is a computer network in which each computer's resources are shared with every other computer in the system. The components of a cluster are usually connected to each other through fast local area network, with each node (computer used as a server) running its own instance of an operating system. Although a cluster may consist of just a few personal computers connected by a simple network, the cluster architecture may also be used to achieve very high levels of performance.

Computer clusters may be configured for different purposes ranging from general purpose business needs such as web-service support, to computation intensive scientific calculations. Fault tolerance allows for scalability, and in high performance situations, low frequency of maintenance routines, resource consolidation and centralized management.

One of the issues in designing a cluster is how tightly coupled the individual nodes may be. For instance, a single computer job may require frequent communication among nodes: this implies that the cluster shares a dedicated network, is densely located, and probably has homogeneous nodes. The other extreme is where a computer job uses one or few nodes, and needs little or no inter node communication, approaching grid computing.

II. Input/Output Techniques

The increase of processing units, speed and computational power, and the complexity of scientific applications that use high performance computing require more efficient Input/Output (I/O) systems. In order to efficiently use the I/O it is necessary to know its performance capacity to determine if it fulfills applications I/O requirements. This proposes a methodology to evaluate I/O performance on computer clusters under different I/O configurations.

In order to evaluate the I/O system performance is necessary know its capacity of storage and throughput. The storage depends on the amount, type and capacity of the devices. The throughput depends on IOPs(Input/Output operations per second) and the latency. Moreover, this capacity is different in each I/O system level. The performance also depends on the connection of the I/O node, the management of I/O devices, placement of I/O node into network topology, buffer/cache state and placement, and availability data and service. Furthermore, to determine if an application uses the whole I/O system capacity, it is necessary to know its I/O behavior and requirements. It is necessary to characterize the behavior of the I/O system and the application to evaluate the I/O system performance. We propose a methodology composed of three phases: Characterization, I/O Configuration Analysis, and Evaluation.

A. Characterization

The characterization phase is divided in two parts: Application and System (I/O system and I/O devices). This is applied to obtain the capacity and performance of the I/O system. Here, we explain the system characterization and the scientific application.

1) I/O System and Devices: Parallel system is characterized at three levels: I/O library, I/O Node(file system) and devices (local file system. We characterize the bandwidth, latency and IOPs for each level.

2) Scientific Application: We have characterized the application to evaluate the I/O system utilization and to know the I/O requirements. The application performance is measured by the I/O time, the transfer rate and IOPs.

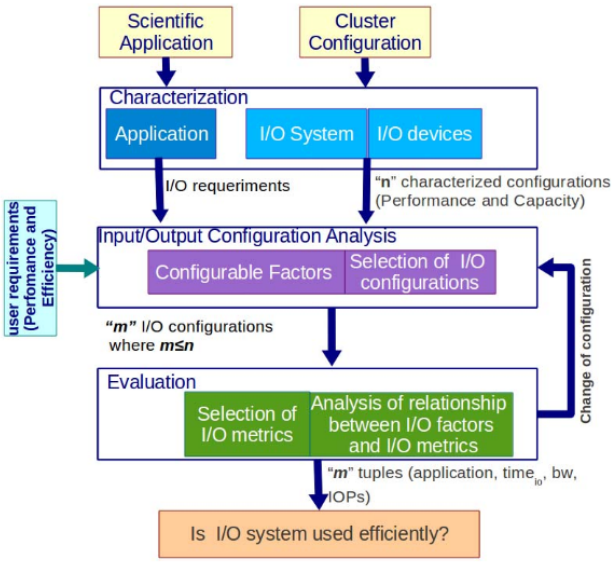

Fig. 1. Methodology for performance evaluation on I/O system

B. Input/Output Configuration Analysis

In this second phase of the methodology, we identify I/O configurable factors and select I/O configurations. This selection depends on user requirements.

1) Configurable Factors: Number and type of file system (local, distributed and parallel), number and type of network (dedicated use and shared with the computing), state and placement of buffer/ cache, number of I/O devices, I/O devices organization (RAID level, JBOD), and number and placement of I/O node.

2) I/O Configuration selection: The configuration is selected based on the performance provided in the I/O path and the RAID level.

C. Evaluation

In the evaluation phase, the application is run on each I/O configuration selected. Application values are compared with characterized values by each configuration to determine the utilization and possible points of inefficiency in the I/O path.

1) Selection of I/O metrics: The metrics for the evaluation are: execution time, I/O time(time to do reading and writing operations), I/O operations per second (IOPs), latency of I/O operations and throughput(number of megabytes transferred per second).

2) Analysis of relationship between I/O factors and I/O metrics: We compare measures of application execution in each configuration with characterized values of I/O path levels.

III. Flow Control Mechanism

The development of a new flow-control mechanism that is able to adjust the buffering resources according to the parallel application communication pattern and the varying activity among communicating peers. In order to show the benefits, we have compared its performance against a static credit-based flow-control mechanism as well as against a communication layer that has unlimited buffering resources, thus not requiring a flow-control protocol.

A. Static Credit-Based Flow-Control

In order to implement the static flow-control version,we have started our development from the commonly used credit-based flow-control mechanism, so that our implementation is a generalization of this mechanism in order to adapt it to our interconnection network.

The original credit-based flow-control mechanism establishes that each sender owns certain buffer space at the receiver's memory. The exact amount of buffering resources is explicitly stated by the number of credits the sender is given at initialization time. In this way, a receiver has as many independent buffers (or buffer partitions) as senders exist. How Notice that a control packet does not consume credits when it is sent, neither it increments the recovered-credit counter when it is extracted from the mailbox. ever, in our interconnection network a receiver only has one buffer, referred to as mailbox, which is shared among all its senders.

A process can simultaneously behave according to two roles: either sender or receiver. We will refer to a process as sender when the process is sending data packets to other processes and it is extracting control packets from its mailbox in order to retrieve credits to continue sending packets. We will refer to a process as receiver when the process is extracting data packets from its mailbox and it is sending back control packets to update the state of its buffers at the sender side.

Initially, the mailbox at each receiver is split into two regions: one for accommodating data packets, data region, and the other one for storing control packets, control region. The data region will be equally distributed among all the processes. Thus, each sender will own an amount of slots in the mailbox of each receiver. This number of credits will be referred to as quota of credits. As each slot can contain one VELO packet, each slot will be equivalent to one credit. In a similar way, the control region will be equally distributed among all the receivers, so that they will have an amount of slots in the mailbox of each sender where to send control packets. The portion corresponding to each receiver will be referred to as control slots. When a receiver reaches the threshold, it sends out a control packet to the sender knowing that, in order for the receiver to reach the threshold, the sender had to free a control slot to obtain more credits to be able to continue sending data packets. This condition ensures that new control packets can be stored at the senders mailbox.

When the initialization stage has finished, the static flow control operation is identical to the original credit-based flow-control mechanism:whenever a data packet is forwarded to the receiver buffer, the credit counter at the sender is decremented. If the counter reaches zero, it means that there is no available slot at the receiver buffer and, therefore,no data packet can be forwarded. On the other hand, when the receiver frees up a slot in its buffer, the recovered-credit count is incremented. When

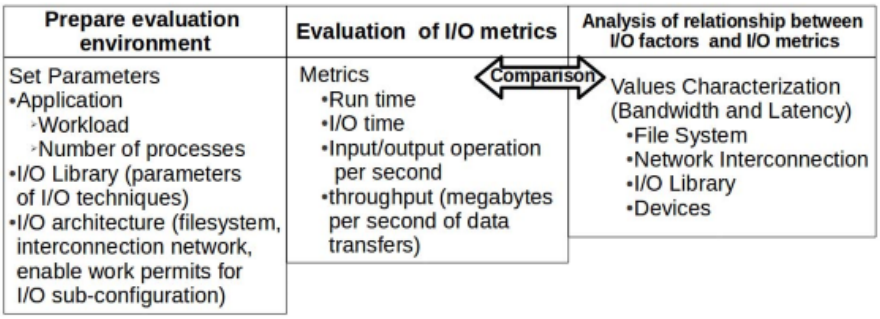

Fig. 2. Evaluation phase

the count reaches the update threshold value, the receiver sends back a control packet to update the sender credit count and then the receiver resets its recovered credit-counter.This type of packets is automatically controlled and the control region allocated in the mailbox.

IV. Storage Architecture

The raw data is written once onto a storage system and then it is read into memory once for analyzing, after which it will seldom be used in the future. Typically, these scientific applications are running on a cluster where the storage system of each node is composed of an array of hard disk drives (HDDs). Although HDDs are economical, they become increasingly incompetent to meet the high I/O performance requirements imposed by these applications. Flash memory based solid-state-drives(SSDs), on the other hand, can provide a high performance and energy-efficiency. Still, they are relatively expensive than HDDs. Here propose a cost-effective yet high performance storage architecture called SOHO (SSD-Workshop-HDD-Warehouse).Its basic idea is to process raw data in the workshop (i.e., SSD), and then, the processed data is moved to the warehouse(i.e., HDD) later.

Batch-processing of big data in real-time or near real-time has steadily become indispensable due to the possibilities given by the emerging hardware techniques and the requirements of scientific research such as seismic wave analysis and gene sequences analysis. However, one of the most challenging is how to efficiently store and process the data with very large sizes. Due to the cost-effectiveness of traditional storage media, HDDs are still the dominant secondary storage devices to offer high capacity for scientific applications. Unfortunately, the performance of big data analyzing is impeded by HDDs because of their low I/O performance.

To better serve the I/O needs of these applications while controlling the cost, we design and implement a novel storage architecture called SOHO (SSD-Workshop HDD-Warehouse) to deliver a near pure-SSD I/O performance without largely increasing the overall cost of storage system in clusters. In the SOHO architecture, an SSD and an HDD are integrated into a hybrid array, which is referred as a SOHO module. The SSD is served as a workshop (or called factory) where raw data

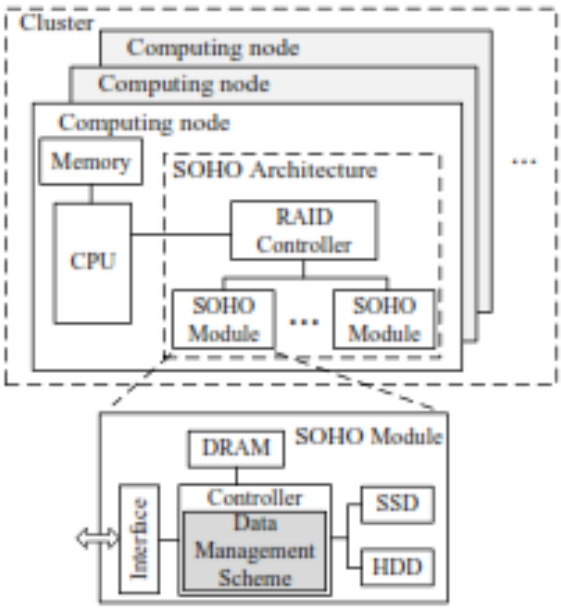

Fig. 3. Architecture of SOHO

is stored and processed, whereas the HDD only offers a huge capacity as a warehouse for storing processed data.

V. Slot Allocation

MapReduce is a popular parallel computing paradigm for large-scale data processing in clusters and data centers. However, the slot utilization can be low, especially when Hadoop Fair Scheduler is used, due to the preallocation of slots among map and reduce tasks, and the order that map tasks followed by reduce tasks in a typical MapReduce environment. To address this problem, we propose to allow slots to be dynamically (re)allocated to either map or reduce tasks depending on their actual requirement. Specifically, we have proposed two types of Dynamic Hadoop Fair Scheduler (DHFS), for two different levels of fairness (i.e., cluster and pool level). In our work, we address the problem of how to improve the utilization and performance of The SOHO architecture has

the following advantages: by utilizing a massive capacity HDD and a moderate size SSD, it noticeable reduces storage system cost in terms of dollar per gigabyte compared with a pure SSD storage system. From the performance point of view, SOHO delivers a very similar I/O performance to that of a pure SSD storage architecture assisted by the intelligent data management scheme.

MapReduce cluster without any prior knowledge or information (e.g., the arriving time of MapReduce jobs, the execution time for map or reduce tasks) about MapReduce jobs. Our solution is novel and straightforward: we break the former first assumption of slot allocation constraint to allow the following:

- Slots are generic and can be used by map and reduce tasks
- Map tasks will prefer to use map slots and likewise reduce tasks prefer to use reduce slots

In other words, when there are insufficient map slots, the map tasks will use up all the map slots and then borrow unused reduce slots. Similarly, reduce tasks can use unallocated map slots if the number of reduce tasks is greater than the number of reduce slots. In this paper, we will focus specifically on Hadoop Fair Scheduler (HFS). This is because the cluster utilization and performance for the whole MapReduce jobs under HFS are much poorer (or more serious) than that under FIFO scheduler.

A. MapReduce

MapReduce is a popular programming model for processing large data sets, initially proposed by Google.Hadoop is an open-source java implementation of MapReduce. When a user submits jobs to the Hadoop cluster, Hadoop system breaks each job into multiple map tasks and reduce tasks. Each map task processes (i.e. scans and records) a data block and produces intermediate results in the form of key-value pairs. Generally, the number of map tasks for a job is determined by input data. There is one map task per data block. The execution time for a map task is determined by the data size of an input block. The reduce tasks consists of shuffle/sort/reduce phases. In the shuffle phase, the reduce tasks fetch the intermediate outputs from each map task. In the sort/reduce phase, the reduce tasks sort intermediate data and then aggregate the intermediate values for each key to produce the final output. The number of reduce tasks for a job is not determined, which depends on the intermediate map outputs.

B. Dynamic Hadoop Fair Scheduler (DHFS)

we first propose two kinds of Dynamic Hadoop Fair Scheduler (DHFS), namely, pool-independent OHFS(PI-DHFS) and pool-dependent OHFS (PO-OHFS).

1) Pool-independent DHFS (PI-DHFS): It considers the dynamic slots allocation from the cluster-level, instead of pool-level. More precisely, it is a typed phase-based dynamic scheduler, i.e., the map tasks have priority in the use of map slots and reduce tasks have priority to reduce slots.

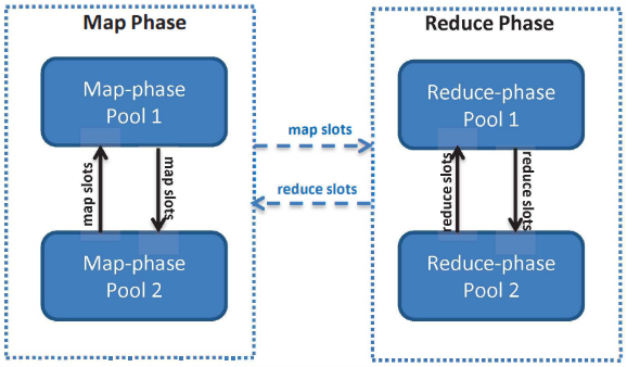

Fig. 4. Example of the fairness-based slots allocation flow for PI-DHFS. The black arrow line and dash line show movement of slots between the map-phase and the reduce-phase pools.

Traditional Hadoop Fair Scheduler (HFS) is a two-level hierarchy. At the first level, HFS allocates task slots across pools, and at the second level, each pool allocates its slots among multiple jobs within its pool. The allocation process consists of two parts, as shown in Fig.4.

- Intra-Phase dynamic slots allocation.
  Each pool is split into two Sub-pools, i.e., map-phase pool and reduce-phase pool. At each phase, each pool will receive its share of slots. An overloaded pool, whose slot demand exceeds its share, can dynamically borrow some unused slots from other pools of the same phase. For example, an overloaded map-phase Pool 1 can borrow map slots from map-phase Pool 2 when Pool 2 is under-utilized, and vice versa.
- Inter-Phase dynamic slots allocation.
  After the intra phase dynamic slots allocation for both the map-phase and reduce-phase, we can now perform dynamic slots allocation across typed phases. That is, when there are some unused reduce slots at the reduce phase and the number of map slots at the map phase is insufficient for map tasks, it will borrow some idle reduce slots for map tasks, to maximize the cluster utilization, and vice versa.

2) Pool-dependent DHFS (PD-DHFS): It is based on the assumption that each pool is selfish, i.e., each pool will always satisfy its own map and reduce tasks with its shared map and reduce slots between its map-phased pool and reduce-phased pool.

It assumes that each pool, consisting of two parts: map-phase pool and reduce-phase pool, is selfish. That is, it always tries to satisfy its own shared map and reduce slots for its own needs at the map-phase and reduce-phase as much as possible before lending them to other pools. PD-DHFS will be done with the following two processes:

- Intra-Pool dynamic slots allocation.
  First, each typed phase pool will receive its share of typed-slots based on max-min fairness at each phase. There are four possible relationships for each pool re-

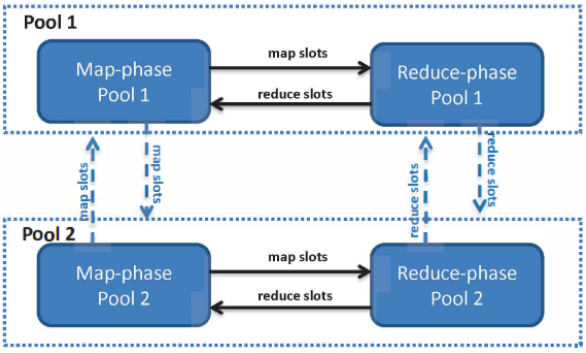

Fig. 5. Example of the fairness-based slots allocation flow for PD-DHFS.The black arrow line and dash line show the borrow flow for slots across pools

garding its demand (denoted as mapSlotsDemand, reduceSlotsDemand) and its share (marked as mapShare, reduceShare) between two phases. We may consider the following cases:

1) Case (a) mapSlotsDemand $<$ mapShare, and reduceSlotsDemand $>$ reduceShare.
   We can borrow some unused map lots for its overloaded reduce tasks from its reduce-phase pool first before yielding to other pools.
2) Case (b) mapSlotsDemand $>$ mapShare, and reduce SlotsDemand $<$ reduceShare.
   In contrast, we can satisfy some unused reduce slots for its map tasks from its map-phase pool first before giving to other pools.
3) Case (c) mapSlotsDemand : mapShare, and reduceSlots Demand: reduceShare.
   Both map slots and reduce slots are enough for its own use. It can lend some unused map slots and reduce slots to other pools.
4) Case (d) mapSlotsDemand $>$ mapShare, and reduce Slots Demand $>$ reduceShare.
   Both map slots and reduce slots for a pool are insufficient. It might need to borrow some unused map or reduce slots from other pools through inter-Pool dynamic slots allocation below.

- *Inter-Pool dynamic slots allocation.*

Obviously we have the following:

1) for a pool, when its mapSlotsDemand + reduceSlotsDemand $<=$ mapShare + reduceShare. The slots are enough for the pool and there is no need to borrow some map or reduceslots from other pools. It is possible for the cases: (a), (b), (c) mentioned above.
2) On the contrary, when mapSlotsDemand + reduceSlots Demand $>$ mapShare + reduceShare, the slots are not enough even after Intra-Pool dynamic slots allocation. It will need to borrow some unused map and reduce slots from other pools, i.e., Inter-Pool dynamic slots allocation, to maximize its own need if possible. It can occurs for pools in the following cases: (a), (b), (d) above.

VI. Task Computing on Clusters

Python is an excellent way to manage Many Task Computing jobs. When the tasks at hand are reliable, and the time they need to execute is predictable, a quick mpi4py implementation is an efficient way to load-balance thousands of tasks quickly and easily. If the runtime characteristics of the tasks in a particular job are unknown or unreliable, then both the IPython Parallel and Celery packages provide a solution. Having fault tolerance for individual processes running on remote cores increases the utility for running task-based parallel problems at a large scale.

VII. Massive Data Computing Framework

MapReduce is a parallel programming technique derived from the functional programming concepts and is proposed by Google for large-scale data processing in a distributed computing environment. The MapReduce model takes a set of key/value pairs as input, and produces a set of key/value pairs as output. The user need to expresses the computation as two functions: Map and Reduce. The MapReduce library groups together all intermediate values associated with the same intermediate key and passes them to the Reduce function. The Reduce function, accepts an intermediate key and a set of values for that key. It merges together these values to form a possibly smaller set of values.

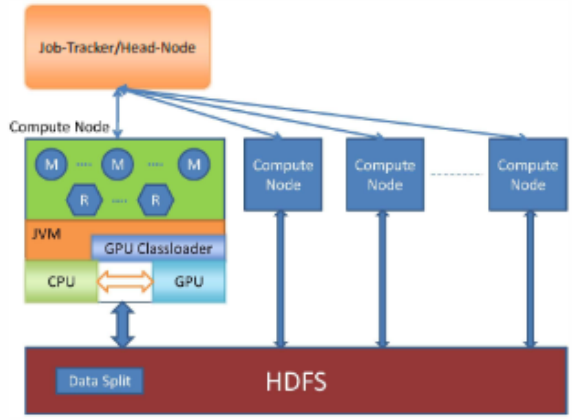

Fig. 6. System Architecture

Our system has two stages, Map and Reduce. Always user should firstly upload the input data to HDFS. HDFS will divide the input data into blocks according to the HDFS configuration. After the user submit the job to the JobTracker, the JobTracker will launch some map tasks on the compute nodes, where the required input blocks located.

In the Map stage, a input split operator divides the input data into multiple splits. Then a RecordReader will read a split and parse the split into key/value pairs. These key/value pairs are the input for the map function. The TaskTracker will start a JVM to run the map function. In our framework, we add some predefined annotations to the source code to indicate some of

the codes will be executed on the GPU. If the map function does not contain any of these annotations, the map function will be executed as a normal Hadoop Map task. Otherwise, the TaskTracker will start a translation process to translate the Java code to CUDA code and compile the CUDA code on-the-f1y. The translation process is managed by a new defined Java class loader, called GPUClassLoader. The GPUClassLoader is also responsible for the memory management and optimization. It will explicitly copy the required data from main memory to the graphics cards memory. After the computation on the GPU, the results will be copied back to the main memory to continue the rest of the map function. After the Map stage is finished, the intermediate key/value pairs will be sorted so that the pairs with the same key are stored consecutively. The Reduce stage of our framework is similarly as the Map stage, if the Reduce function is annotated, it will call the GPU to do the computation, otherwise, it will act as a normal Hadoop Reduce function.

VIII. CONCLUSION

A methodology to analyze I/O performance of parallel computers has been proposed and applied. As future work, we aim to define an I/O model of the application to support the evaluation, design and selection of the configurations. The development of a new flow-control mechanism that is able to adjust the buffering resources according to the parallel application communication pattern and the varying activity among communicating peers. In order to show the benefits, we have compared its performance against a static credit-based flow-control mechanism.

Write-once-read-once scientific applications running on clusters normally process huge amounts of data in a batch manner, which demands a high performance storage system for each computing node. They generally split the I/O workloads between SSD and HDD, which is not suitable for write-once-read-once scientific applications. This is because HDDs are incompetent to provide a sufficient I/O performance.

Dynamic Hadoop Fair Schedulers (DHFS) to improve the utilization and performance of MapReduce clusters while guaranteeing the fairness. The core technique is dynamically allocating map (or reduce) slots to map and reduce tasks. Two types of DHFS are presented, namely, PI-DHFS and PD-DHFS, based on fairness for cluster and pools, respectively. we have optimized the data flow to eliminate the unnecessary data transfers between CPU memory and GPU memory. This can effectively improve the performance of our framework. Meanwhile, we have also observed that if some data transfer instructions can be combined into one instruction, the data transfer will be efficiently reduced.

REFERENCES

[1] Sandra Mendez, Dolores Rexachs and Emilio Luque, “Methodology for Performance Evaluation of the Input/Output System on Computer Clusters”, 2011, IEEE International Conference on Cluster Computing.

[2] Javier Prades, Federico Silla, Jose Duato Holger Froning, Mondrian Nussle, “A New End-to-End Flow-Control Mechanism for High Performance Computing Clusters”, 2012, IEEE International Conference on Cluster Computing.

[3] Cailiang Xu, Wei Wang, Deng Zhou, and Tao Xie,“An SSD-HDD Integrated Storage Architecture for Write-Once-Read-Once Applications on Clusters”, 2015, IEEE International Conference on Cluster Computing.

[4] Shanjiang Tang, Bu-Sung Lee, Bingsheng He, “Dynamic Slot Allocation Technique for MapReduce Clusters”, 2013 IEEE International Conference on Cluster Computing (CLUSTER).

[5] Monte Lunacek Jazcek Braden Thomas Hauser , “The Scaling of Many-Task Computing Approaches in Python on Cluster Supercomputers”, 2013 IEEE International Conference on Cluster Computing (CLUSTER).

[6] Yanlong Zhai, Emmanuel Mbarushimana, Wei Li, Jing Zhang, Ying Guo, “A High Performance Massive Data Computing Framework Based on CPU/GPU Cluster”, Science and Technology on Complex Systems Simulation Laboratory, Beijing, China.

Opportunities and Challenges of Electronic Payment Systems

Jithinkrishna I V, Risheen E A and Treesa Soyet Joy
Vidya Academy of Science & Technology
Thrissur - 680501, India

Dijesh P
Associate Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur - 680501, India
(email: dijesh.p@vidyaacdemy.ac.in)

***Abstract*—In this paper an overview of electronic payment methods and systems is given. E-commerce provides the capability of buying and selling products, information and services on the internet and other online environments. In an e-commerce environment, payment take the form of money exchange in an electronic form, and are therefore called electronic payment. Today India is at a stage of demonetization so; in the present scenario this study is inevitable to makes electronic payments at any time through the internet directly to manage the e-business environment. This study aimed to identify the issues and challenges of electronic payment systems and offer some solutions to improve the e-payment system. Security is the protection of e-commerce assets from unauthorized access, use, alteration, or destruction.**

***Index Terms*—E-commerce, e-payment system, e-commerce security**

I. Introduction

THE ELECTRONIC payment system is considered as the backbone of e-commerce and one of its most crucial aspects. It can be defined as a payment service that utilizes the Information and communication technologies including integrated circuit (IC) card, cryptography, and telecommunication networks' (Raja et. al., 2008). An efficient electronic payment system lessens the cost of trading and is thought to be essential for the functioning of capital and inter-bank markets. With the advancement of technology, electronic payment system has taken many forms including credit cards, debit cards, electronic cash and check systems, smart cards, digital wallets contactless payment methods and mobile payments and so on.

E-commerce has become a rapidly growing market today. With the proliferation of tablets and smartphones, the use of electronic payment methods has grown up to 21% in 2012 (Rau, 2013). The use of credit cards was the major international means of online payment that dominated in a variety of transaction markets. It was estimated that 95% of all e-commerce transactions in the United States are performed using credit cards(Abrazhevich, 2004). Other widely used online payment alternatives are debit cards (with rising number of users worldwide) and online payment systems like Paypal, Stripe or Skrill. With the availability of a variety of electronic payment means including mobile payments, mediating services, and electronic currency, an appropriate option can be chosen for a particular type of transaction (Paunov and Vickery, 2006).

The future of a specific electronic payment system depends upon how it overcomes the practical and analytical challenges faced by various means of online payments. These challenges include issues of law and regulation (buyer and seller protection), technological capabilities of e-payment service providers, commercial relationships, and security considerations such as verification and authentication issues (Paunov and Vickery, 2006.

II. Major Online Payment Systems

Studying various systems of electronic payments, Koponen (2006) explained that there are a wide variety of online payment systems that have been developed in past few years and these systems can be broadly classified into account-based and electronic currency systems. Account-based systems allow users to make payments via their personal bank accounts; whereas the other system allows the payment only if the consumer possesses an adequate amount of electronic currency. These systems offer a number of payment methods that include:

- Electronic payment cards (debit,credit, and charge cards)
- E-wallets
- Virtual credit cards
- Mobile payments
- Loyalty and Smart cards
- Electronic cash (E-cash)
- Stored-value card payments

Paunov and Vickery (2006) gives a description of electronic payment methods in their report evaluating the online payment systems for e-commerce, a summary of this description is given here to look at various characteristic features of the most commonly used online payment services.

Requirements	Transaction Process	Limitation/Risk
1. Online Account in Digital Wallet. For example few popular e wallets are PayUmoney, Paytm, Pockets, Oxygen wallet, Mobiwiki etc. 2. A mobile phone with wallet app loaded. Bank wallets are also used operated through Desktop PC. 3. Internet connection.	1. Download wallet app in mobile and create account using mobile no. Mobile number is treated as wallet account no. 2. Load money using debit/credit card or net banking. 3. Link your bank account with digital wallet to transfer money in advance to your wallet. 4. Transfer money from one wallet to other using mobile number.	1. Consumer Wallet Limits: Rs.20, 000/month for all. Rs.1 lakh/month with KYC 2. Merchant Wallet Limits: Rs.50, 000/month with Self Declaration. Rs.1 lakh/month with KYC 3. Money can be transferred to same company wallet.

Fig. 1. E-Wallets

A. Credit Cards

The most commonly used online payment mode so far was the use of credit cards. Initially, the security concerns hindered in the adoption of credit cards for making online payments but later with the provision of more secure features to protect every transaction made, customers developed trust on the use of credit cards. Applicability of credit cards is a strong factor that contributed to its wide use throughout the world.

Credit card companies have established a wide network for their consumers ensuring a huge user base for a number of different transactions. However, it is considered a less-suitable method for small businesses and customers that need to make small payments due to high fees for credit cards (Paunov and Vickery, 2006). Aggregation or cumulative payment solution can be a way to adapt credit card payment system for micro-payments. One of the major advantages of credit cards is their easy to use functionality with making online transactions in no time and from anywhere. These cards are easy to obtain and use as customers dont need to purchase any extra software or hardware to work with them. Cardholder authentication procedure is also simple, with the provision of a name, credit card number, and expiry date. For the security of consumers' personal information, credit card companies have developed a number of complementary systems including MasterCard SecureCode and Verified by Visa. These systems allow users to create a password and use it when they shop online through their credit cards.

B. Debit Cards

The popularity of the debit cards is constantly rising and currently debit cards the most popular non-cash payments instrument globally (Capgemini and RBS, 2013). In contrast to credit cards, payments through debit cards are withdrawn directly from the personal account of the consumer instead of an intermediary account. This makes it difficult for consumers to handle payment disputes as there funds don't have an extra protection in a debit account. For debit payments, providing the account number is enough without the necessity of producing a physical card or card number. The use of debit cards is particularly high in most countries with a specific user base depending on the conditions and regulations attached to the issuance of credit cards. However, debit payments may not popular on merchant websites as debit cards do not cater the demand for payments made by international customers (Paunov and Vickery, 2006). Since there are lower costs for using debit cards unlike credit cards this method is suitable for micropayments. In addition, the overall security of debit card payments is found to be higher than that of credit card payments with extensive identification requirements demanded by the banks.

C. Mobile Payments

According to Hoofnagle, et al.(2012), payments made through wireless devices like mobile phones and smartphones are thought to provide more convenience, reduce the fee for the transaction, and increase the security of electronic payment. This payment system has also made it easier for businesses to collect useful information about their customers and their purchases. Paunov and Vickery(2006) found the applicability of mobile payment systems to be quite wide due to the remarkable growth and greater penetration of mobile devices as compared to other telecommunication infrastructure.

Requirements	**Transaction Process**	**Limitation/Risk**
1. Issue of Debit Cards by Bank. Card Pin/password or mobile for OTP (One Type Password) verification. 2. Bank ATMs 3. Swipe machine or POS (Point of Sale) machine at merchant. 4. Online payment portal	1. Bank issue ATM card with a PIN no. 2. Used to withdraw cash from any ATM machine using PIN no. 3. Used at any POS for shopping. Also for online shopping 4. SMS notifications come in mobile for every transaction.	1. POS of Swipe machine at merchant is a must. 2. Cloning of card is a security threat.

Fig. 2. Unified Payments Interface (UPI)

Mobile payment methods are suitable for offline micropayments as well as for online purchases. This method is a potential attraction for online traders due to an enormous user base of mobile phones. The use of mobile payment service does not only reduce the overall cost of a transaction but also offer a better payment security. However, mobile payment systems have encountered certain challenges in obtaining a significant consumer base for a number of reasons including privacy issues and their inability to cater international payments.

D. Mobile Wallets

In a study regarding consumer adoption of mobile wallets, Doan (2014) explained that Mobile wallet is formed when your Smartphone functions as a leather wallet: it can have digital coupons, digital money(transactions), digital cards, and digital receipts. Mobile wallet service allows the user to install an application from online stores in their smartphones and use them to pay for their online and offline purchases. Using latest technologies that connect smartphones to the physical world such as NFC (Near Field Communication), sound waves, and QR codes, cloud-based solutions, mobile wallets are believed to provide more convenient payment solutions to the customers in future(Husson, 2015).

E. Electronic Cash

During initial stages of introducing online payment systems, electronic cash systems proposed in the form of DigiCash or CyberCash. However, these systems were not much appreciated and disappeared soon. At present, smart card-based systems are more common in use for the payment of small amounts by many businesses. Smart cards usually rely on specific hardware and card reader for their use and authentication. In addition to smart cards, numerous electronic cash systems have also been established such as Virtual BBVA and Clic-e. These systems work with the use of prepaid cards or electronic tokens that represent a certain value and can be exchanged for hard cash (Paunov and Vickery,2006).

III. Security

Viruses are a nuisance threat in the e-commerce world. They only disrupt e-commerce operations and should be classified as a Denial of Service (DoS) tool. The Trojan horse remote control programs and their commercial equivalents are the most serious threat to e commerce. Trojan horse programs allow data integrity and fraud attacks to originate from a seemingly valid client system and can be extremely difficult to resolve. A hacker could initiate fraudulent orders from a victim system and the ecommerce server wouldnt know the order was fake or real. Password protection, encrypted client server communication, public private key encryption schemes are all negated by the simple fact that the Trojan horse program allows the hacker to see all clear-text before it gets encrypted.

A. Purpose of security

1) Data confidentiality is provided by encryption/decryption.
2) Authentication and identification ensures that someone is who he or she claims to be is implemented with digital signatures.
3) Access control governs what resources a user may access on the system. Uses valid IDs and passwords.
4) Data integrity ensures info has not been tampered with. Is implemented by message digest or hashing.
5) Non-repudiation is intended not to deny a sale or purchase Implemented with digital signatures.

A cryptographic algorithm is called a cipher. It is a mathematical function. Most attacks are focused on finding the key.

B. Security Issues

E-commerce security is the protection of e-commerce assets from unauthorized access, use, alteration, or destruction. While security features do not guarantee a secure system, they are necessary to build a secure system.Security features have four categories:

1) Authentication:

Requirements	**Transaction Process**	**Limitation/Risk**
1. Issue of Debit Cards by Bank. Card Pin/password or mobile for OTP (One Type Password) verification. 2. Bank ATMs 3. Swipe machine or POS (Point of Sale) machine at merchant. 4. Online payment portal	1. Bank issue ATM card with a PIN no. 2. Used to withdraw cash from any ATM machine using PIN no. 3. Used at any POS for shopping. Also for online shopping 4. SMS notifications come in mobile for every transaction.	1. POS of Swipe machine at merchant is a must. 2. Cloning of card is a security threat.

Fig. 3. Debit/ATM Cards

Requirements	**Transaction Process**	**Limitation/Risk**
1. Issue of Credit Cards by Bank. Card Pin/password or mobile for OTP (One Type Password) verification. 2. Swipe machine or POS (Point of Sale) machine at merchant. 3. Online payment portal	1. Bank issue Credit cards with a PIN no to only eligible customer. 2. There is credit limit for issued cards, limit vary from person to person depending upon income. 3. Used at any POS for shopping. Also for online shopping or transaction. 4. Every month Bill is generated, the total dues is to be paid before the due date, otherwise interest is charged. 5. SMS notifications come in mobile for every transaction.	1. Every card has a credit limit, beyond that you cannot shop. 2. Cash withdrawal is possible but at huge interest rate. 3. POS of Swipe machine at merchant is a must. 4. Cloning of card is a security threat.

Fig. 4. Credit Cards

Verifies who you say you are. It enforces that you are the only one allowed to logon to your Internet banking account.

2) Authorization:
Allows only you to manipulate your resources in specific ways. This prevents you from increasing the balance of your account or deleting a bill.
3) Encryption:
Deals with information hiding. It ensures you cannot spy on others during Internet banking transactions.
4) Auditing: Keeps a record of operations. Merchants use auditing to prove that you bought a specific merchandise.
5) Integrity:
Prevention against unauthorized data modification
6) Non-repudiation:
Prevention against any one party from reneging on an agreement after the fact
7) Availability:
Prevention against data delays or removal.

C. Security Threats

There are three types of security threats: Denial of service, unauthorized access, and theft and fraud.

1) *Security(DOS)*
Two primary types of DOS attacks: spamming and viruses.
- Spamming:

Sending unsolicited commercial emails to individuals E-mail bombing caused by a hacker targeting one computer or network, and sending thousands of email messages to it. Surfing involves hackers placing software agents onto a thirdparty system and setting it off to send requests to an intended target. DDOS (distributed denial of service attacks) involveshackers placing software agents onto a number of third-party systems and setting them off to simultaneously send requests to an intended target

- Viruses:
  Self-replicating computer programs designed to perform unwanted events.

The methods used for unauthorised access are the following:

- Worms:
  Special viruses that spread using direct Internet connections.
- Trojan horses:
  Disguised as legitimate software and trick users into running the program.

2) *Security (unauthorized access)*
   a) Illegal access to systems, applications or data
   b) Passive unauthorized access:
      - Listening to communications channel for finding secret
      - May use content for damaging purposes
   c) Active unauthorized access
      - Modifying system or data
      - Message stream modification
   d) Changes intent of messages, e.g., to abort or delay a negotiation on a contract
   e) Masquerading or spoofing sending a message that appears to be from someone else.
      - Impersonating another user
      - Name (changing the from field) or IP levels (changing the source and/or destination IP address of packets in the network)
   f) Sniffers: software that illegally access data traversing across the network.
   g) Software and operating systems security holes
3) Security (theft and fraud)
   - Data theft already discussed under the unauthorized access section
   - Fraud occurs when the stolen data is used or modified.
   - Theft of software via illegal copying from companys servers.
   - Theft of hardware, specifically laptops.

IV. Conclusion

Electronic payment refers to the mode of payment which does not include physical cash or cheques. It includes debit card, credit card, smart card, e-wallet etc. E-commerce has its main link in its development on line in the use of payment methods, some of which we have analyzed in this work .The risk to the online payments are theft of payments data, personal data and fraudulent rejection on the part of customers. Therefore, and until the use of electronic signatures is wide spread, we must use the technology available for the moment to guarantee a reasonable minimum level of security on the network.

References

[1] Bhasker, Bharat (2013), *Electronic Commerce, Framework, Technologies and Applications*, McGraw Hill Education (India) Private Limited, p.9.2-9.16.
[2] Madan, Sushila (2013), *E-Commerce*, Mayur Paperbacks. P.4.4-4.35.
[3] S.S.Hugar, *Trends and challeges to Indian Banking*, Deep & Deep publications, New Delhi.
[4] Kaur Manjot, *E-Commerce*, Kalyani Publcation, New Delhi
[5] Abrazhevich, Dennis (2004), *Electronic Payment Systems: a User-Centered Perspective and Interaction Design.* Netherlands: Technische Universiteit Eindhoven
[6] Aigbe, Princewill and Akpojaro, Jackson (2014), " Analysis of Security Issues in Electronic Payment Systems", Nigeria: International Journal of Computer Applications (0975-8887), Vol. 108.
[7] Jean Lassignardie, Kevin Brown (2013), *World Payments Report 2013*, Capgemini and The Royal Bank of Scotland.
[8] Doan, Ngoc (2014), *Consumer adoption in Mobile Wallet*, The Turku University of Applied Sciences.
[9] Hoofnagle, Chris Jay, Urban, Jennifer M. and Su Li (2012), "Mobile Payments: Consumer benefits and new privacy concerns", Berkley Center for Law and Technology (BCLT) Research Paper.
[10] Husson, Thomas (2015), "The Future of Mobile Wallets lies beyond Payments", U.S.A.: Forrester Research Inc.
[11] ISACA (2011), *Mobile Payments: Risk, Security and Assurance Issues.* U.S.A.: ISACA Emerging Technology White Paper.
[12] Houssam El Ismaili1, Hanane Houmani, Hicham Madroumi, "A Secure Electronic Transaction Payment Protocol Design and Implementation", Morocco: International Journal of Advanced Computer Science and Applications, Vol.5, No.5.

Cyber Security Warning Systems

John Francis, Neha Pauly and Sradha K S
Vidya Academy of Science & Technology
Thrissur- 680501, India

Manesh D
Assistant Professor of Computer Applications
Vidya Academy of Science & Technology
Thrissur- 680501, India
(email: manesh.d@vidyaacademy.ac.in)

***Abstract*—The cyber security in one of most important security today that was much needed in the world of internet. In this paper we analysed how we can categorise the cyber users via social-technical theories of information systems security and social-technical approach for cyber security warning systems. Then we discuss about the problems that arrived during past and present problems. In the next we discuss about the risk analysis via Bayesian which identifies the threats that going to attack the cyber network. After that we model and generate a cyber situation awareness. Then by using lucent technology we identifies attacks and by CSRM Methodology we solve that threats.**

***Index Terms*—Cyber security, social-technical approaches to warning systems, threats to networks, lucent technology, CSRM methodology**

I. Introduction

CYBER attacks, cyber crimes etc are the common issues faced in the cyber world. These are happening due to the insecurity in the cyber world. This paper describes risk analysis in cyber situation awareness using Bayesian approach, socio technical approach for cyber security warning systems and a security risk management methodology designed to ensure that assurance and resilience are built into mission system as they progress through the acquisition and system development life cycle and continuing during system operation. Bayesian network classifier is a method to analyze the network traffic in the cyber network. It helps to solve the uncertainty of cyber network and the currently existing cyber security warning systems dont take into account the different security culture of security warnings recipients. To address this vulnerability we suggest developing a platform that close or eliminate this existing socio-technical gap between cyber security warning disseminators and end recipients. This in turn would improve the effectiveness of the taken measures in response to these warnings.

The suggested platform name is: A Socio-Technical Cyber Security Coordination System ST(CS)2. Its role is to provide a collaborative platform between member organizations (subscribers) and cyber security warning systems in cyber security risk management. The methodology also identifies how well existing policies and requirements counter the identified treats to mission systems. The methodology is based primarily on the NIST risk management standards and is consistent with current DOD policy, which requires the qualitative amassment and subsequent prioritization of cyber security risk. At the conclusion we understand that Bayesian network classifier can analyze DoS (Denial of Service) with the help of association rule mining and genetic algorithm.

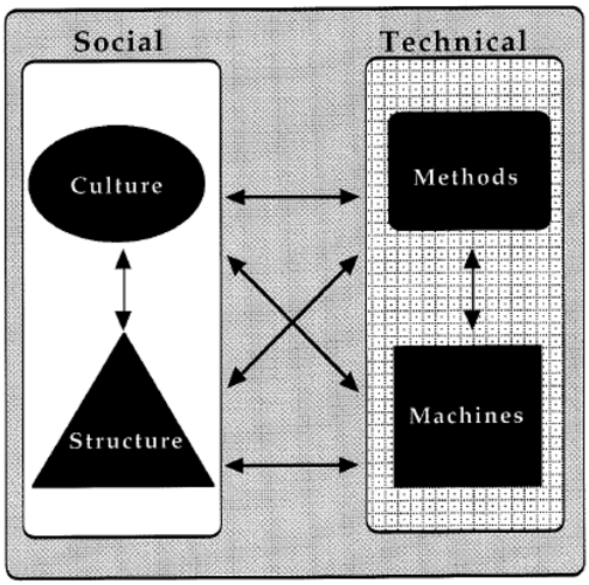

Fig. 1. A socio-technical system

II. Socio-technical Theories of Information Systems Security

A. Culture theory of risk

One of the early researches on the social nature of information systems security is the "cultural theory of risk. In this theory people were characterized into four groups:

- Individualistic:
  The individualistic group comprises those having benign perception of their surrounding world. The group subscribers are known to be risk seekers.
- Egalitarian:
  Egalitarians have ephemeral perception of things surrounding them. The subscribers of this group are known to be risk averse.

- Hierarchical
  Hierarchical group subscribers have preserve/tolerant perception of their world and are willing to accept risks within limits.
- Fatalistic:
  Lastly, fatalists group subscribers have capricious perception of things around them in general. They are the risk neutral.

B. Socio-technical theory

In the 1990s, Kowalski proposed the theory of socio-technical systems. He has argued that every socio-technical system is affected by four components belonging to two subsystems:

- Culture and structure belonging to the social subsystem
- Methods and machines belonging to the technical subsystem

As the system needs to maintain a state of equilibrium, any change happen to one of the system components due to an internal or external factor, the system other components need to interchange accordingly to keep maintaining the state of equilibrium of the whole system.

Based on the socio-technical systems theory, Kowalski has further developed the "Security by Consensus (SBC) framework to compare different national approaches to information security . The framework divide the security problem into social and technical areas:

- Ethical/cultural
- Legal/contractual
- Managerial/administrational
- Operational/procedural
- Technical

Technical area is further divided to:

- Application
- Operating System
- Hardware

Kowalski argues that the framework is required to ensure secure communication between different socio-technical systems.

The Security Continuum or the Security Value Chain is another suggested socio-technical framework for modeling security culture of certain entity. The chain is composed of the main five security access control types:

- Deterrent:
  Deterrent controls are used to discourage attackers from committing an abuse or violating an existing security policy.
- Protective/Preventive:
  Preventive controls are implemented to avoid the occurrence of a security incident e.g. firewalls.
- Detective
  Detective controls are used in order to detect unwanted incident after it occur.
- Corrective/Respond

Fig. 2. The SBC model

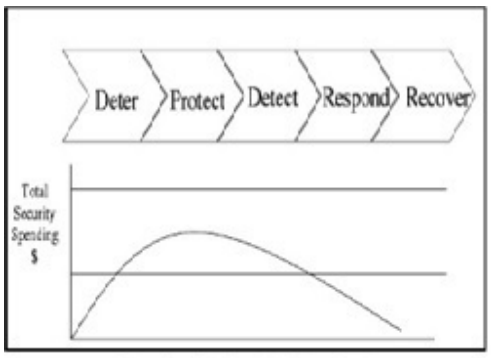

Figure 3. The Security Value Chain

Fig. 3. The security value chain

  Corrective controls are used to remediate a situation after the occurrence of a security event.
- Recovery
  While recovery controls are implemented to restore an asset after the security incident, it more takes a monetary form.

III. Proposal for Socio-technical Approach for Cyber Security warning Systems

Given the fact that cyber security is a global problem that requires collaboration and coordination between member countries. There is a need to develop a global cyber security culture including laws and policies related to cyber security practices. However, the fact we discussed earlier about the existence of different perceptions and understanding of security risks and countermeasures, different laws and legislations and different policies still constitute a major obstacle that hinders such development. Moreover, and to our best knowledge and research, the currently existing cyber security warning systems

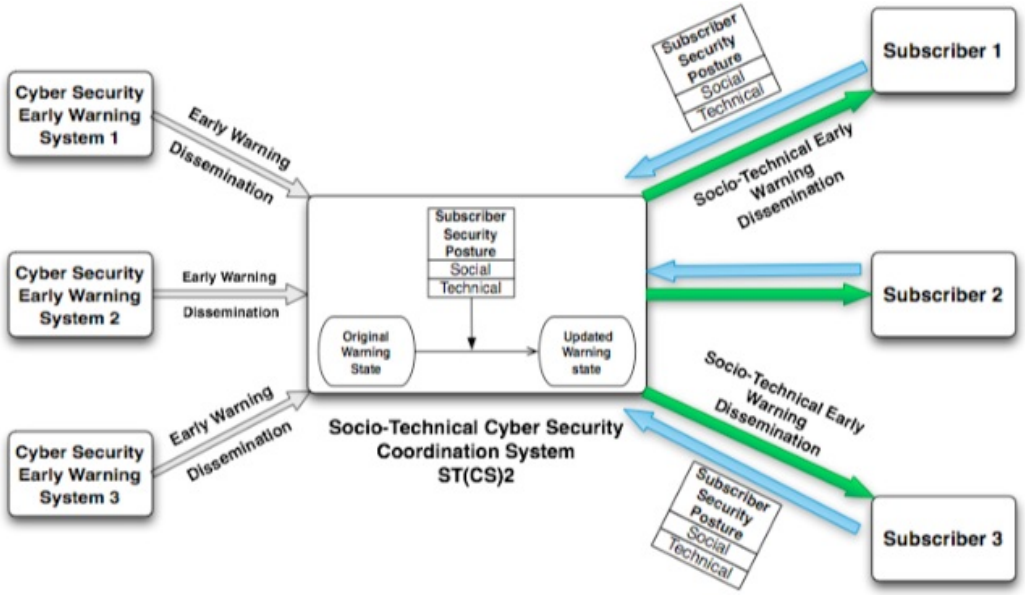

Fig. 4. The socio-technical cyber security coordination system ST(CS)^2

dont take into account the different security culture of security warnings recipients.

To address this vulnerability we suggest developing a platform that close or eliminate this existing socio-technical gap between cyber security warning disseminators and end recipients. This in turn would improve the effectiveness of the taken measures in response to these warnings. The suggested platform name is: A Socio-Technical Cyber Security Coordination System ST(CS)^2. Its role is to provide a collaborative platform between member organizations (subscribers) and cyber security warning systems. The platform operators who are socio-technical security experts should ensure that subscribers will receive guided cyber security warnings based on each subscriber socio-technical security posture.

The idea of the collaborative feature was inspired form the ITU-IMPACT global response center NEWS and ESCAPE platforms which together provide collaborative platform for domain experts and cyber security warning systems with the exception that they dont tackle the social aspect of security. The difference between general and guided cyber security warnings is that general warnings give broad overview of the current situation. Also most of the time their data sources are not indicated . In guided warnings the threat level and recommended countermeasures are customized based on the socio-technical security posture of the warning recipient. This customization is made upon social and technical factors related to the existing gap between the cyber security warning required countermeasures and the implemented security measures at the warning recipient site. These implemented measures reflect how security is understood. Also the existing social and technical security gaps determine the alert level for the guided warning. As subscribers to the ST(CS)^{2}2 platform, member organizations have to regularly feed the platform operators with information about their security implementations at the different social and technical areas e.g. policies, operations, practices, technical implementations. These areas correspond to the different layers of the SBC framework discussed earlier. The platform operators will then analyze this information to get an understanding of each subscribing organization security culture. Also the social and technical vulnerabilities will be identified. Of course the analysis process should be based on number of socio-technical security metrics. For this purpose we suggest using the security values chain as a metric to identify existing security gaps at social and technical levels.

When new cyber security warning system is disseminated by a cyber security warning system, the ST(CS)^2 operators will collect this warning and analyze it from socio-technical perspective. These required countermeasures are then compared to each subscriber implemented security controls. The existing gap between the required countermeasures and the subscriber security social and technical posture is then identified. Accordingly the warning threat level is determined and required actions are identified. The ST(CS)2 platform will then disseminate a guided version of the warning to the subscriber. Guided warning will ensure those subscribes are effectively responding to security warnings.

IV. Warning System Problems: Past and Present

A. Warning Systems in General

Regardless if it is a warning system for bad weather or for cyber attacks there are a number of basic and common design principles and functional requirements that are needed to be considered when proposing any alarm/alert and warning system. Four such principles have been outlined by the Norwegian Petroleum Directorate and state that the system must display information that is relevant to the stakeholders' roles

at the time, indicate what response is required, be presented at a rate that the stakeholders can deal with, and be easy to understand by all stakeholders. Sir Francis Beaufort, a British admiral, introduced around 1800 a standardized way of describing the weather conditions. This is today known as the Beaufort scale. It defines 17 (originally 12) categories of wind strength, based on objective observations of the environment. Examples are the behavior of smoke or damages to trees. Another well know system is the (Modified) Mercalli intensity scale, which was developed around 1900. It classifies the strength of earthquakes into 12 groups.

A number of different attempts to establish warning systems that have been proposed use different types of security metrics. These proposals can be grouped into two:

1) Generic cyber threat levels for the public
2) Specialized metrics for a specific situation

B. Generic Warning Systems.

The first group aims to give a broad overview of the current environment, usually focusing on consumers. Examples are the indices that are published by anti-virus vendors . They summarize their current view of virus and other malware activity. Although the result is a categorization, they often do not publish the details of the metric. Only results from the same vendor can therefore be compared. Some governments also publish their view of the current cyber threat level . While some of the investigated metrics do have a sufficient description, none of their data sources is stated. This makes the quality of the information questionable. Lastly, some well known internet organizations, such as the Internet Storm Center , also publish a status based on their data collection. But again, in case of the ISC no sources are documented.

C. Specialized Warning Systems.

These metrics in this group aim to provide information about a specific environment. Their aim is therefore to aid specialists in their task. All the investigated systems were described in theory. No practical, ongoing implementation was found to be publicy available. On a lower technical level, defines a system to predict future attacks through analysis of existing data. Variable length Markov models are used together with the assumption that an attacker follows a certain sequential logic. The results show that the intended prediction is possible.

The MITRE Corporation suggested the Cyber Prep framework . It defines threat levels for an organization based on the intention and capabilities of the potential attacker. This allows for a targeted and effective response to the suspected threats.

D. Review of the Current Situation.

The preceding review of some of the research in the area of cyber weather forecasting show very well the issues at hand. Currently, there exists no functional framework to assess the cyber threat level in an objective/quantitative manner. Those in the first group are too generic, and the algorithm and data basis behind them are not documented sufficiently. Those in the second group aim to provide protection for organizations, but the predictions are made for a very short term for technical countermeasures or they are very generic and independent of the dynamic environment.

V. Review of Bayesian Inference

Many aspects of cognition can be formulated as solutions to problems of induction. Given some observed data about the world, the mind draws conclusions about the underlying process or structure that gave rise to these data, and then uses that knowledge to make predictive judgments about new cases. Bayesian inference is a rational engine for solving such problems within a probabilistic framework, and consequently is the heart of most probabilistic models.

A. The Bayesian Model

Bayesian model grows out of a formula known as Bayes rule (Nave Bayes). When stated in terms of abstract random variables, Bayes rule is no more than a result of probability theory. Assume two random variables, A and B. One of the principles of probability theory (sometimes called the chain rule) allows joint probability to be written of these two variables taking on particular values a and b, $P(a, b)$, as the product of the conditional probability that A will take on value a given B takes on value b, $P(a|b)$, and the marginal probability that B takes on value b, $P(b)$. Thus, it can be written as

$$P(a, b) = P(a|b)P(b).$$

Using factorization with the choice of B rather than A, the joint probability is written as

$$P(a, b) = P(b|a)P(a).$$

From above we have

$$P(a|b)P(b) = P(b|a)P(a).$$

Thus

$$P(a|b) = P(a|b)P(b)/p(a).$$

This expression is Bayes rule, it indicates the computation of the conditional probability of b given a, from the conditional probability of a given b. Bayes rule gets its strength and weakness based on some assumptions with respect to the variables under consideration.

B. The Bayesian Model for Cyber Security

The following illustrates some of the characteristics of the Bayesian model for cyber security.

Consider an attack on a system in a network, the system may become at risk of any form of attack as a result of the use of network resources, an event represented by the variable Denial of Service (DoS) attack (denoted by D). Such an attack can cause damage to systems or lead to denial of service, an event represented by the variable teardrop (denoted by TD). The DoS attack might result from status flag of connection, represented by the variable sf (denoted by S) or connection protocol, represented by the variable `http` (denoted by H).

It is reasonable to assume that a network user is at risk of a remote to local ($R2L$) attack, an event represented by the variable Imap (denoted by I). All variables representation are binary; thus, they are either true (denoted by T) or false (denoted by F). The condition probability table (CPT) of each node is listed besides the node.

In this illustration the parents of the variable D are the nodes S and H. The child of D is TD, and the parent of I is $R2L$. Following the Bayesian Network (BN) independence assumption, several independence statements can be observed in this case. For example, the variables H and S are marginally independent, but when D is given they are conditionally dependent. This relation is often called explaining away.

When H is given, S and D, are conditionally independent. When D is given, TD is conditionally independent of its ancestors H and S. The conditional independence statement of the BN provides a compact factorization of the Joint Probability Distributions (JPD). Instead of factorizing the joint distribution of all the variables by the chain rule:

$$P(S,H,I,D,TD) = P(S)P(H|S)P(I|H,S)P(D|I,H,S)P(TD|D,I,H,S)$$

The BN defines a unique JPD in a factored form, i.e.

$$P(S,H,I,D,TD) = P(S)P(H)P(I|S)P(P|H,S)P(TD|D)$$

Note that the BN form reduces the number of the model parameters, which belong to a multinomial distribution in this case, from to 25-1=31 to 10 parameters. Such a reduction provides great benefits from inference, learning (parameter estimation), and computational perspective. The resulting model is more robust with respect to bias-variance effects. A practical graphical criterion that helps to investigate the structure of the JPD modelled by a BN is called d-separation. It captures both the conditional independence and dependence relations that are implied by the Markov condition on the random variables.

C. Inference via Bayesian Network

Given a BN that specified the JPD in a factored form, one can evaluate all possible inference queries by marginalization, i.e. summing out over irrelevant variables. Two types of inference support are often considered: predictive support for node Xi, based on evidence nodes connected to Xi through its parent nodes (also called top-down reasoning), and diagnostic support for node Xi, based on evidence nodes connected to Xi through its children nodes (also called bottom-up reasoning).

The belief for the occurrence of a status flag given the observation that the network suffers from the risk of teardrop. A connection protocol is represented by H, which is one of the network protocol used for network communication and the status flag is represented by S. The probability that the network will suffer from a DoS attack such as Teardrop (TD) is defined by the occurrence (true or T value) of S and H. The values are chosen randomly to demonstrate the dependency.

D. Types of Attack

The attacks fall into four main categories:

1) Denial of Service (DoS): This is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine such as smurf, neptune;
2) Remote to Local Attack (R2L): This occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user on that machine such as guess password, phf;
3) User to Root Attack (U2R): This is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system such as buffer overflow, perl;
4) Probing Attack: This is an attempt to scan a computer or a network of computers to gather information for the apparent purpose of circumventing its security controls such as portsweep, nmap.

Classification probability for attack represents the probability of occurrence of each attack type for a given data set.

$$PA = \frac{\sum TA}{\sum n},$$

where n is the total number of dataset, TA is the total number of occurrence of a type of attack in the same dataset and PA is the probability of classification of the attack type.

VI. Framework for Cyber Security Situation Awareness

The framework for network security situation awareness proposed is based upon of two parts, the modelling of network security situation and the generation of network security situation. Actually the modelling of network security situation is the construction of a model for measuring the network security situation based upon the D-S Theory. And the generation of network security situation consists of three phases 1-acquiring attack pattern, 2-transforming the discovered frequent patterns and sequential patterns to the correlation rules of alert events, 3-implementing the dynamically generation of network security situation graph.

A. The Modelling of Network Security Situation

The modelling of network security situation is used for to construct the standardized data model suited for the measuring of network security situation, and support the general process of the simplification, filtering, and fusion of alert events from security situation sensors. The objective of event simplification is to merge the redundant alert events of the identical attack detected from several sensors. The objective of event filtering is to remove those events dissatisfied with the constraint requirements, and these constraint requirements are stored in the knowledge base in the form of attribute or rule according

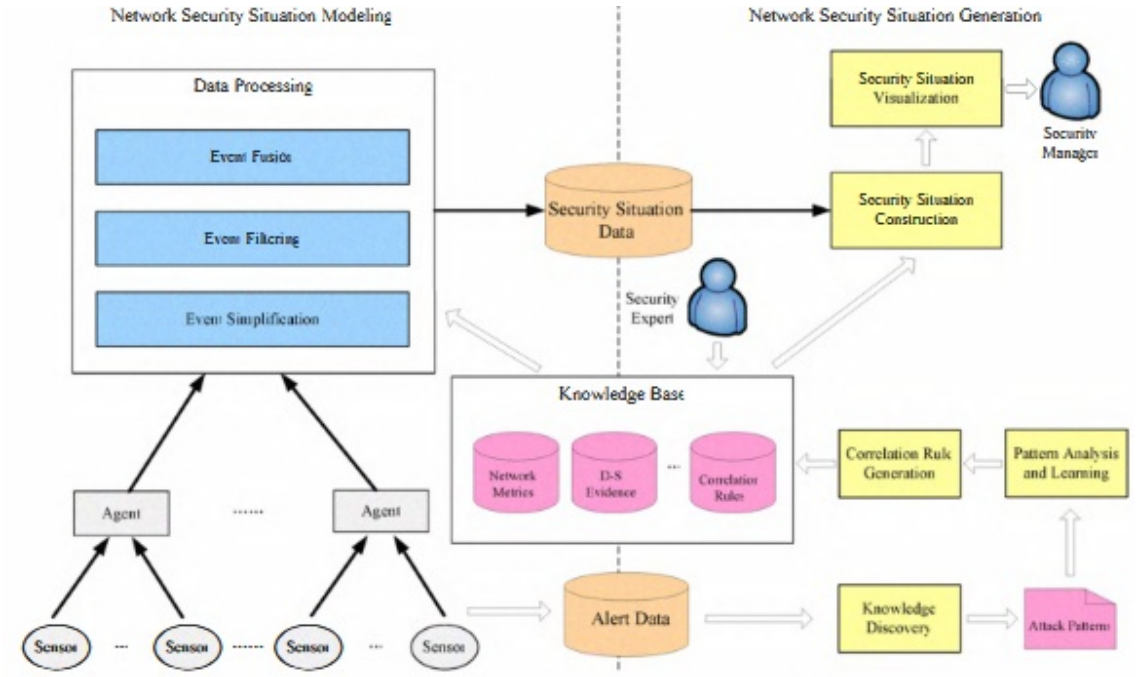

Fig. 5. Framework for cyber security awareness

to the requirements of network security situation awareness. For instance, the events can be removed if the key attributes of these events are absent or out of the required ranges, because they are meaningless for the analysis of network security situation. Through the processing of simplification and filtering, the repeated security events are merged, the amount of security events is greatly reduced and the abstraction degree is improved, at the same time the security situation information implied are preserved.

The objective of event fusion is to introduce the different confidence level to the security events that received from different sensors and have already pre-processed, simplified and filtered, quantitatively evaluate these security events by fusing multiple attributes, so as to effectively reduce the false positive and false negative of security alerts and provide support for the inference, analysis and generation of network security situations. The foundation of event fusion function is Dempster Sharer (D-S) evidence theory.

B. The Generation of Network Security Situation

The generation of network security situation. The two network security situation data sources available for knowledge discovery: one is the set of security alert events generated from the attack simulations; the other is the set of historical security alert events.

The generation of network Security consists of three phases.

1) Simplification and filtering of security alert events:
   In this phase the distributions of various types of security events are statistically analysed via automatic tools; secondly, the meaningless events are deleted by evaluating the importance of each type of alert events based upon the rules of simplification and filtering, which uses D-S evidence theory as the foundation of event processing.
2) Knowledge discovery from the set of security alert events:
   In this phase we use three algorithms of frequent pattern and sequential pattern for the knowledge discovery. The frequent pattern refers to the correlations among the attributes of events, and the objective of which is to infer the constraints among the attributes of events and transform to the filtering rules after adding correlation actions. The three main algorithms are:
   - Frequent Pattern Mining:
     The most significant feature of frequent pattern mining algorithm (such as, FP-Tree algorithm) is to compress the large database to the compact tree structure (FP-tree) and quickly mine the set of frequency patterns without the need of generating the candidate items, since it avoids the repeated database scanning.
   - Sequential Pattern Mining:
     WINEPI algorithm is used in sequential pattern mining to discover the sequential relationship among security alert events.
   - Pattern Analysis and Learning:
     To utilize the discovered knowledge effectively, Prolog-EBG machine learning algorithm are adopted to properly interpret and analyse the discovered knowledge by introducing the prior knowledge of the domain, and the revised and optimized rules are exported from the set of security alert events generated from the attack simulations.

3) The last phase of the generation of network security situation is to generate network security situation refers to the correlation of network security events, construction of network security situation graph, and assessment of the global network security situation. Net-SSA periodically update the network security situation graph based upon the security situation data calculated from event fusion and correlation.

The network security situation awareness system we developed by ourselves was applied in the experiment. This system includes a network security situation generation engine based on knowledge discovery. The test data LLDOS was provided by MIT Lincoln Lab, which was collected under the attack inspect situation of DARPA2000. LLDOS 1.0 was the first data collection which was created by DARPA. It was consists of five attack stages:

1) Get the list of active hosts
2) Find weak Solaris hosts
3) Invade the system by Solaris Sadmind buffer overflow bug
4) Install mstream DDoS Trojan on the controlled hosts
5) Start attack on the remote server by the controlled hosts.

According with the security situation modeling, alert events generated from various security sensors were simplified, filtered, fused and correlated. The number of the warning events decreased greatly from 64481 to 6164. At the same time, according to the correlation rule, it converts many trivial attacks which aimed at the victim host from Forged IP into a DDoS attack.

At the end we analysed the existing problems of network security situation awareness and proposed a framework based on knowledge discovery can be reduced by the modeling of network security situation and generation of network security situation The framework consists of the modeling of network security situation and the whole process of the generation of network security situation. We have described the construction of the formal model for network security situation measurement based upon the D-S evidence theory, the extraction the frequent patterns and sequential patterns from the dataset of network security situation based upon knowledge discovery method and the transformation of these patterns to the correlation rules of network security situation, and the automatic generation of network security situation graph. Application of the integrated Network Security Situation Awareness system (Net-SSA) shows that the proposed framework supports for the accurate modeling and effective generation of network security situation.

Network and computers often hold a companys most precious and costly commodities. if the network and computer vulnerabilities of any large network such as Lucent technologies or mission systems are not identified and mitigated they could enable an intruder to seriously compromise the security of companies network and data.

VII. The Case of Lucent Technologies Inc.

Lucent Technologies has a data network of more than 100,000 computers, referred to in this paper as the lucent intranet. The security of this data network is of paramount importance. However, the task of managing its security is also a monumental challenge. To safeguard the Lucent intranet against threats, we must perform risk assessments to effectively manage all aspects of security, which can be categorized as cyber security and non-cyber security .The solutions to each aspect of securityy may appear to differ in their details, they all share a common objective: to minimize the risks to an asset by recognizing its internal and external threats, identify-in its potential vulnerabilities, performing risk assessmints, and then mitigating the vulnerabilities accordingly. Here security is categorized into networking security and os security networking security concerns the security of the software that controls and manages how one computer communicates with other computers Operating system security concerns the security of the software that controls and manages how a computer processes user operations and programs.

At present, there are at least 100 known network-in vulnerabilities in the TCP/IP environment and at least 100 known operating system vulnerabilities associated with each of the UNIX or Windows NT* plat-forms. About 50 networking vulnerabilities and about 25 operating system vulnerabilities are considered critical, because by compromising them an intruder could attain the privileged state of that computer and render the computer defenceless. The software tool we used to probe networking security recognizes 48 vulnerabilities as high risk, including those that provide authorized access to a computer and the associated network. Similarly, the software tool used to probe operating system security recognizes 20 vulnerabilities in UNIX and 26 vulnerabilities in Windows NT as required to change; these vulnerabilities pose an immediate threat to the security of a computer. To simplify the terminology, we use the term critical to denote both the terms of high risk in networking vulnerability and required to change in operating system vulnerability. Each computer can be considered as operating in two states privilege and the user state. The combinations of some networking and operating system vulnerabilities could result in an intruders attaining the privileged state.

The traditional approach to cyber security is to identify all vulnerabilities, assess their risks, and then mitigate them on a vulnerability-by-vulnerability and computer-by-computer basis. Obviously, it is not oper-ationally feasible or desirable to take this approach to the lucent intranet. We decided to try an innovative approach by combining the statistical sampling and analysis methodology with the networking and operating system security discipline.

In mid-1997, the security people initiated a cyber security profile project and developed a sampling plan based on the assumption that vulnerabilities follow a Poisson distribution; they have also collected data for analysis and validated the methodology. They found that the distribution-of networking

Fig. 6. CSRM processes

vulnerabilities and operating system vulnerabilities across all computers are highly concentrated in a few areas. In addition, they determined that the root causes of these selected vulnerabilities fall into two broad categories. These findings led them to formulate a phased Lucent-wide security strategy that focuses on these frequently occurring, critical vulnerabilities and applies a mitigation plan that addresses the two root causes. As a result, they succeeded in approaching the problem of managing cyber security within the Lucent intranet by developing not only an effective means of identifying vulnerabilities, but also an efficient means of mitigating them. If properly deployed, this ability will be to the 21st century what quality control charts were to the last.

A. Sampling Plan and Data Collection

The Security concerns are analysed in two phases: Phase I for networking security networking vulnerability are the weakness associated with a computer that occurs because the computer is part of a network of computers. The vulnerabilities are analysed by using poison distribution and these are categorized into: Configuration errors and software bugs.

B. Operating Security Profile

In Phase 2 the identification and mitigation of vulnerabilities associated with operating systems such as unix and windowsNT are concerned and these vulnerabilities are analysed and grouped into four broad categories.

- RTC
- RSH
- INF
- N/A

VIII. CSRM Methodology

CSRM was developed to provide decision-makers with detailed knowledge and background information that will enable them to manage Information System (IS)-related security risk. Unlike an audit or investigative report, which focuses on discrepancies, CSRM provides an objective, systematic, and analytical approach to assessing system security risk to enable senior management to better understand system risks and allocate resources to reduce and correct potential losses and operational impacts.

CSRM starts with the NIST methodology, which encompasses three processes: risk assessment, risk mitigation, and evaluation and assessment, and expands it to include key concepts from programmatic risk management, such as risk management planning and risk monitoring and control.

IX. Conclusion

Computer security is the protection of computer system from theft or damage to their hardware, software or electronic data as well as from disruption or misdirection of the services they provide. The security aspect of computer can be categorized into cyber security and non-cyber security. These strategies or solution to each aspect of security may appear to differ in their details. Our aim was to develop strategies to minimize security threats to our system.

In this short paper we outline the problems associated with cyber security and socio-technical cyber security. Technical area is further divided to application, operating system and hardware. The Security Value Chain is another suggested socio-technical framework for modeling security culture of certain entity. As mentioned the main security problems in cyber world are associated with network security and operating security. Network security problems can be identified using Bayesian approach and by using CSRM methodology. In BAYESIAN approach the security concerns are identified and treated by using Bayes' theorem in probability. Where as in CSRM approach it is treated and identified by using Poisson distribution in probability. The modelling of network security situation is the construction of a model for measuring the network security situation based upon the D-S Theory. And the generation of network security. situation consists of three phases. By using CSRM methodology threat is identified and categorised into two broad classifications and then it is mitigated based on the risk it provide to the system.

On other hand operating system security threats are those security problems that are associated with operating systems such as windowsNT and unix. To identify the risk ,here HSS and NSS scans are performed and then the risks are categorised into four main types namely RTC,RSH,INF,N/A and based on this categories the risks are mitigated.

Another important concept which was discussed in this paper is about creating a security warning system. The most difficult challenge, and a major problem for alarm/alert systems is to design a system that is easy for all stakeholders to understand. A number of different attempts to establish warning systems that have been proposed use different types of security metrics.

References

[1] InfoSec Institute, "DoS Attack and Free DoS Attacking Tools, (Online) Available: `http://resources.infosecinstitute.com/dos-attacks-free-dos-attacking-tools`, 2013.

[2] R. Lipschutz, "After Authority: War, Peace and Global Politics in the 21st Century Albany: State University of New York Press, 2000.

[3] The World Bank Group, Cyber Security: A New Model for Protecting theNetwork. (Online) Available `http://siteresources.worldbank.org/Resources/CyberSecurity.pdf`, 2005.

[4] S.L. Yang, S. Byers, J. Holsopple, B. Argau and D. Fava, "Intrusion Activity Projection for Cyber Situational Awareness, Rochester Institute of Technology, 2009.

[5] B. K. Alese, A. J. Gabriel, and A. O. Adetunmbi, "Design and Implementation of Internet Protocol Security Filtering Rules in a Network Environment, International Journal of Computer Science and Information Security (IJCSIS), vol. 9, no. 7, July 2011.

[6] KDD Cup 99 (Online) Available: `http://kdd.ics.uci.edu/databases/kddcup99/task.html`, 1999.
[7] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems, In Human Factors Journal, vol 37(1), pp 32-64, 1995.
[8] R. Baldwin, "Rule Based Analysis of Computer Security, Technical Report TR-401, MIT LCS Lab, 1989.
[9] D. Farmer, and H. Eugene, "The COPS Security Checker System, Technical Report CSD-TR-993, Purdue University, 1999.
[10] P. Ammann, W. Duminda,, and K. Saket, "Scalable, Graph-based Network Vulnerability Analysis, In Proceedings of 9th ACM Conference on Computer and Communications Security, Washington, DC, 2002.
[11] S. Cheung, "Modeling Multistep Cyber-attacks for Scenario Recognition, In DARPA Information Survivability Conference and Exposition (DISCEXIII), 2003.

Name Index